An artificial intelligence platform for the diagnosis and surgical planning of strabismus using corneal light-reflection photos

Keli Mao[1#], Yahan Yang[1#], Chong Guo[1#], Yi Zhu[2], Chuan Chen[3], Jingchang Chen[1], Li Liu[4], Lifei Chen[1], Zijun Mo[5], Bingsen Lin[5], Xinliang Zhang[5], Sijin Li[5], Xiaoming Lin[1*], Haotian Lin[1*]

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China; [2]Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Miami, FL, USA; [3]Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, USA; [4]The Shunde Hospital, Southern Medical University (The First People's Hospital of Shunde), Foshan, China; [5]Zhongshan Medical School, Sun Yat-sen University, Guangzhou, China

Contributions: (I) Conception and design: H Lin, X Lin; (II) Administrative support: H Lin; (III) Provision of study materials or patients: H Lin, X Lin; (IV) Collection and assembly of data: K Mao, Y Yang, L Liu, L Chen, Z Mo, B Lin, X Zhang, S Li; (V) Data analysis and interpretation: K Mao, Y Yang, C Guo, C Chen, Y Zhu, J Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

*These authors contributed equally to this work as co-senior authors.

Correspondence to: Haotian Lin, MD, PhD; Xiaoming Lin, MD. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, 54 Xianlie Road, Guangzhou 510060, China. Email: gddlht@aliyun.com; linxiaom@mail.sysu.edu.cn.

Background: Strabismus affects approximately 0.8–6.8% of the world's population and can lead to abnormal visual function. However, Strabismus screening and measurement are laborious and require professional training. This study aimed to develop an artificial intelligence (AI) platform based on corneal light-reflection photos for the diagnosis of strabismus and to provide preoperative advice.

Methods: An AI platform consisting of three deep learning (DL) systems for strabismus diagnosis, angle evaluation, and operation plannings based on corneal light-reflection photos was trained and retrospectively validated using a retrospective development data set obtained between Jan 1, 2014, and Dec 31, 2018. Corneal light-reflection photos were collected to train the DL systems for strabismus screening and deviation evaluations in the horizontal strabismus while concatenated images (each composed of two photos representing different gaze states) were procured to train the DL system for operative advice regarding exotropia. The AI platform was further prospectively validated using a prospective development data set captured between Sep 1, 2019, and Jun 10, 2020.

Results: In total, 5,797 and 571 photos were included in the retrospective and prospectively development data sets, respectively. In the retrospective test sets, the screening system detected strabismus with a sensitivity of 99.1% [95% confidence interval (95% CI), 98.1–99.7%], a specificity of 98.3% (95% CI, 94.6–99.5%), and an AUC of 0.998 (95% CI, 0.993–1.000, $P<0.001$). Compared to the angle measured by the perimeter arc, the deviation evaluation system achieved a level of accuracy of ±6.6° (95% LoA) with a small bias of 1.0°. Compared to the real design, the operation advice system provided advice regarding the target angle within ±5.5° (95% LoA). Regarding strabismus in the prospective test set, the AUC was 0.980. The platform achieved a level of accuracy of ±7.0° (95% LoA) in the deviation evaluation and ±6.1° (95% LoA) in the target angle suggestion.

Conclusions: The AI platform based on corneal light-reflection photos can provide reliable references for strabismus diagnosis, angle evaluation, and surgical plannings.

Keywords: Artificial intelligence (AI); machine learning; strabismus; corneal light-reflection photos

Submitted Jul 22, 2020. Accepted for publication Dec 18, 2020.
doi: 10.21037/atm-20-5442
View this article at: http://dx.doi.org/10.21037/atm-20-5442

© Annals of Translational Medicine. All rights reserved.

Introduction

Strabismus, characterized by binocular misalignment, affects approximately 0.8–6.8% of the world's population and appears by the age of 3 years in 65% of affected individuals (1-5). As a leading cause of impaired binocular vision and abnormal visual function, strabismus compromises the quality of life of preschool children (6). Timely establishment of binocular alignment can improve long-term vision and sensorimotor outcomes (7-10). Thus, patients with symptomatic misalignment or deviation greater than 12 prism diopters (PD) require extraocular muscle surgery if eyeglasses and amblyopia management fail to align the eyes, and the accurate measurement of the deviation is the foundation of medical intervention

Manual measurement of deviation is often laborious and highly dependent on the experience of the specialist and the cooperation of the patients. The alternate prism cover test (APCT), which is the gold standard for angle measurement, is time-consuming, and the interexaminer variation is reported to be approximately 10 PD (11,12). In addition, the perimeter arc method, where strabismus is measured by moving a flashlight along the perimeter arc until the light reflects in the center of the pupil of the deflected eye, also requires patient cooperation and is therefore difficult to perform in young children (13-15). Other methods, including the Hirschberg and Krimsky tests, are substantially less accurate even when carried out by experienced strabismologists (16).

Previous studies have attempted to develop reliable systems based on digital photographs or videos to facilitate automatic diagnosis and evaluation. As widely adopted designs, 2D or 3D eye models combined with feature extraction are heavily dependent on fixed parameters of model eyes (17,18), which significantly limits their application in patients with large deviation or high refractive error. Additionally, the thickness of video-oculography (VOG) goggles restricts their application in young children; accordingly, the accessibility represents another obstacle to the implementation of methods that are based on special devices such as infrared cameras (19,20).

Recently, artificial intelligence (AI) has gradually changed healthcare. As a branch of computer science, AI aims to create intelligent machines that are able to perform tasks by mimicking human intelligence, such as visual perception and voice recognition. With recent progress in AI based on deep learning (DL), Major advances in diagnostic technologies are offering unprecedented insight into ocular disease such as cataract, glaucoma and diabetic retinopathy (21). Using representation-learning methods with multiple levels of abstraction, the DL system automatically recognizes visual patterns without the need for manual feature engineering based on training on large datasets of labeled images (22). Given the strong performance of DL in identifying ocular diseases based on varieties of images such as slit lamp photography and ultra-widefield fundus images (22-24), the development of DL presents an unprecedented opportunity to provide a reliable reference for strabismus diagnosis, evaluation, and surgical planning.

In this study, we developed an AI platform based on corneal light-reflection photos to facilitate the diagnosis and angle evaluation of strabismus and to provide advice for surgical planning.

We present the following article in accordance with the Materials Design Analysis Reporting (MDAR) checklist and the STROBE reporting checklist (available at http://dx.doi.org/10.21037/atm-20-5442).

Methods

A diagnostic, cross-sectional study.

Datasets

Corneal light-reflection photos

This study followed the tenets set forth in the Declaration of Helsinki (as revised in 2013), and approval was obtained from the Ethics Committee of the Zhongshan Ophthalmic Center of Sun Yat-sen University (2019KYPJ153). Written informed consent was obtained from the participants or their legal guardians. For training and retrospective testing on the first stage, all images were taken with a Nikon D5300 (Nikon Co., Tokyo, Japan). Patients were asked to sit facing forward on an examination chair approximately 33 cm away from the camera and stare at an accommodative picture attached over the camera's objective lens. To obtain a corneal light-reflection image, a steady, soft point of light was placed next to the camera rather than using the camera flash, which emits harsh light and causes some patients to close their eyes. The photographer ensured that at least one of the patient's eyes was looking forward without face tilt. Two photos were taken of each patient with alternate strabismus, one with the right eye staring straight and the other with the left. Photos were taken with and without glasses in patients with corrective lenses. The resolution of each photograph was 2,922×2,900 pixels. For prospective

© Annals of Translational Medicine. All rights reserved.

testing in the second stage, images were taken with three different devices, including a Nikon D5200 (Nikon Co., Tokyo, Japan), a Nikon D5300 (Nikon Co., Tokyo, Japan) and an iPhone 8 (Apple Inc., California, USA) by three photographers in the same manner.

Data sources

In the first stage, data were obtained from two datasets, including a hospital-based dataset and a population-based dataset. The hospital-based dataset, which included photos obtained from patients before and after surgery between Jan 1, 2014, and Dec 31, 2018, was derived from the Zhongshan Ophthalmic Center (ZOC) in China. The deviations were measured based on the perimeter arc, a method widely used in China, which has an accuracy of 1 degree of angle-error-free (13-15). Both the strabismus measurements and the operative designs were performed and recorded by professional ophthalmologists. The population-based dataset included a combination of photos obtained from students of a middle-high school in Guangdong Province as part of a physical examination and photos obtained from officials at our department from May 1st, 2018, to Jun 1st, 2018. The subjects included in the dataset were examined by a professional ophthalmologist. The examination consisted of two parts, the strabological evaluation, which was completed with a 4△ test through an alternate cover test and a sensorial evaluation performed using a Titmus stereoscopic acuity test. If the results of these two evaluations were normal, the patient was defined as orthotropic; otherwise, the patient was defined as strabismic (25).

All of these photos were used for the establishment and retrospective testing of the screening system. Photos of horizontal strabismus patients with accurate records of preoperative deviation were used for the establishment and testing of the deviation evaluation system. The operation advice model relied on photos of exotropia patients (except for those with paralytic strabismus or large variation (>10 PD) (6) in the strabismus angle across different fixation states) who had undergone successful initial surgeries.

In the second stage, the data were obtained from patients who visited the outpatient clinics of the ZOC between Sep 1, 2019, and Jun 10, 2020 (NCT04416776). Among these participants, deviation measurements and operative designs were further performed and recorded for patients who received an operation. The enrollment criteria for prospective testing of each system as well as the ophthalmologist team responsible for examination and operation were consistent with those in the first stage.

Image preparation

The image preparation process is shown in *Figure 1*. For the screening system, the quality of the photos was first evaluated by a group of ophthalmologists. The exclusion criteria for the screening system were as follows: photos with noise points, or with covered reflex points caused by blepharoptosis. Then, among the eligible items, the quality of photos obtained in horizontal strabismus patients was further assessed for the deviation evaluation system. The exclusion criteria included photos with blurred reflex points, reflex points outside the edge of the cornea, the presence of vertical strabismus, or an absence of full deviation for intermittent exotropia.

The photos were then concatenated horizontally to include states in which either eye was staring forward in patients with alternate exotropia or two identical photos obtained from a patient with monocular exotropia, such as patients with perceptual exotropia, for the operation advice system. In this step, patients with refractive error provided images obtained with glasses, while in other patients, images were obtained without glasses. One stitched image was labeled as eligible only if the photos obtained on both sides were judged eligible according to the criteria used in the deviation evaluation system.

A similar image preparation process for prospective testing was conducted for prospective testing in the second stage (*Figure 2*).

Development of the algorithm

The AI platform contained three DL models as shown in *Figure 3*.

Screening system

The dataset containing a total of 5,797 images were randomly divided into training, validation, and retrospective test sets at ratios of 70%, 15%, and 15%, respectively.

The convolution neural network (CNN) architecture InceptionResNetV2 (26) was used in this study. Weights were initialized using those previously pretrained on the ImageNet dataset. The output of the last convolution layer was connected to a global average pooling layer, then a 512-neuron hidden layer with ReLU activation, then a dropout layer with a probability of 0.5, and finally to a one-neuron output layer with a sigmoid activation function. The loss function used was binary cross-entropy. The ADAptive Momentum (ADAM) optimizer was applied with an initial learning rate of 0.001, beta 1 of 0.9, beta 2 of 0.999, fuzz

© Annals of Translational Medicine. All rights reserved.

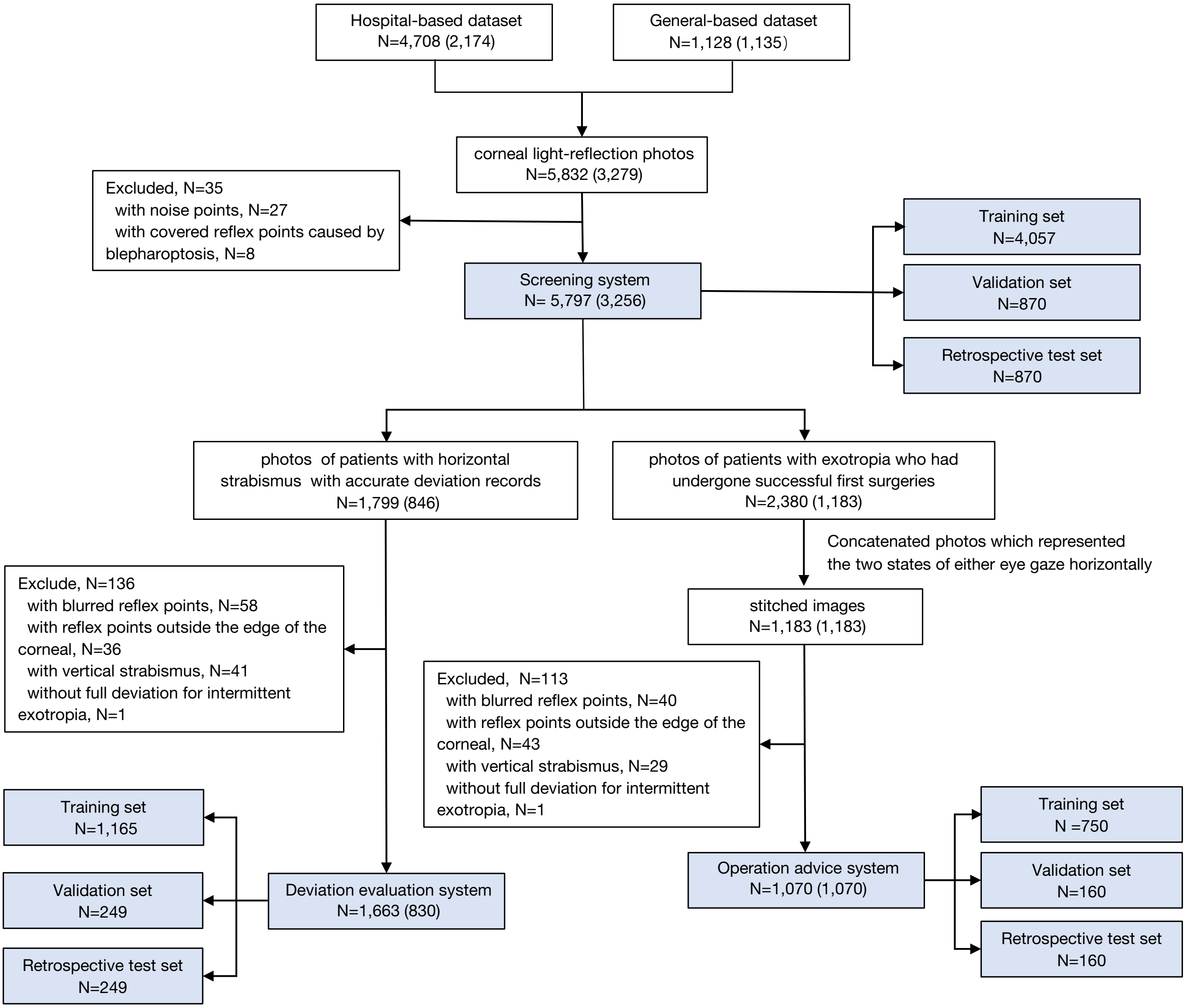

Hospital-based dataset
N=4,708 (2,174)
General-based dataset
N=1,128 (1,135)
corneal light-reflection photos
N=5,832 (3,279)
Excluded, N=35
with noise points, N=27
with covered reflex points caused by blepharoptosis, N=8
Screening system
N= 5,797 (3,256)
Training set
N=4,057
Validation set
N=870
Retrospective test set
N=870
photos of patients with horizontal strabismus with accurate deviation records
N=1,799 (846)
photos of patients with exotropia who had undergone successful first surgeries
N=2,380 (1,183)
Concatenated photos which represented the two states of either eye gaze horizontally
stitched images
N=1,183 (1,183)
Exclude, N=136
with blurred reflex points, N=58
with reflex points outside the edge of the corneal, N=36
with vertical strabismus, N=41
without full deviation for intermittent exotropia, N=1
Excluded, N=113
with blurred reflex points, N=40
with reflex points outside the edge of the corneal, N=43
with vertical strabismus, N=29
without full deviation for intermittent exotropia, N=1
Training set
N=1,165
Validation set
N=249
Retrospective test set
N=249
Deviation evaluation system
N=1,663 (830)
Operation advice system
N=1,070 (1,070)
Training set
N =750
Validation set
N=160
Retrospective test set
N=160

Figure 1 Data collection, image preparation, and group-splitting processes used in the first stage for training and retrospective testing. The numbers of participants are indicated in brackets.

factor of 1e-7, and zero learning rate decay.

The images were resized to 128 pixels high by 512 pixels wide, and the pixel values were normalized to between 0 and 1 before the images were input into the CNN. Data augmentation, including horizontal mirror flipping, clockwise and counter-clockwise rotation up to 5 degrees, horizontal and vertical shift up to 1/20 of the height and width of the image, and brightness shift between 0.6 and 1.4 of the original pixel intensity, was performed on the training set to increase the sample size and the model's generalizability to unseen data.

A batch size of 10 was used. The model was trained up to 500 epochs, with early stopping applied. If the validation loss did not decrease for 120 consecutive epochs, the training process would terminate early. The model state in which the validation loss was the lowest during training was recorded as the final model state.

Deviation evaluation system

The dataset contained 1,663 samples, of which 1,065 (70%) belonged to the training set and 249 (15%) belonged to each of the validation and retrospective testing sets.

© Annals of Translational Medicine. All rights reserved.

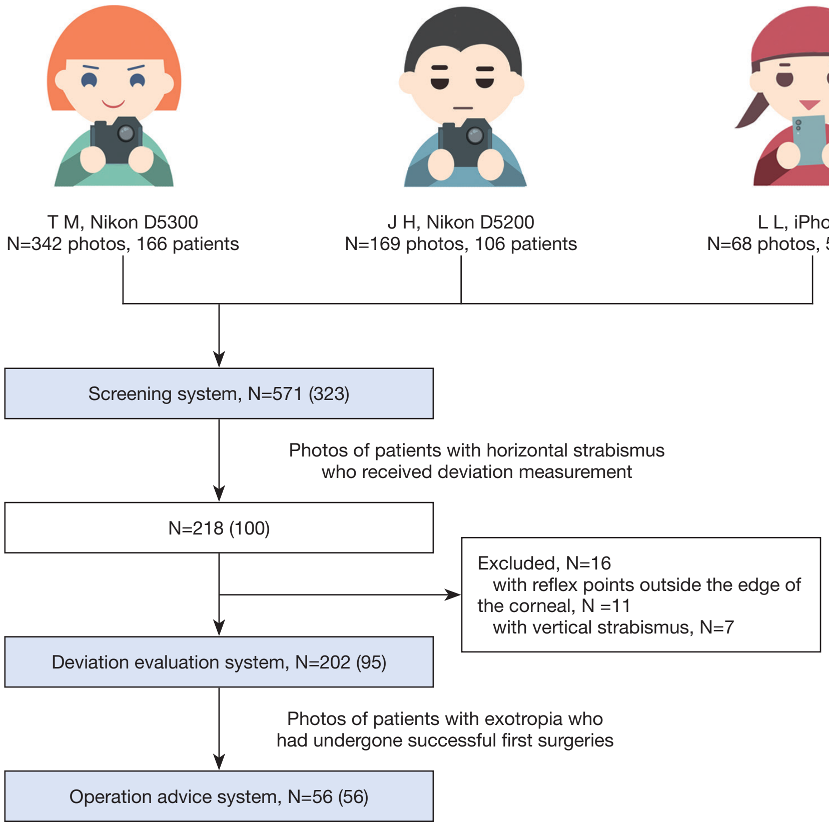

T M, Nikon D5300
N=342 photos, 166 patients
J H, Nikon D5200
N=169 photos, 106 patients
L L, iPhone 8
N=68 photos, 58 patients
Screening system, N=571 (323)
Photos of patients with horizontal strabismus who received deviation measurement
N=218 (100)
Excluded, N=16
with reflex points outside the edge of the corneal, N =11
with vertical strabismus, N=7
Deviation evaluation system, N=202 (95)
Photos of patients with exotropia who had undergone successful first surgeries
Operation advice system, N=56 (56)

Figure 2 Data collection and image preparation used in the three systems in the second stage for prospective testing. The numbers of participants are indicated in brackets.

The same CNN architecture as the one used in the classification task was employed, with the activation function for the last output neuron changed from sigmoid to linear. The loss function used was the mean square error (MSE). The mean average error (MAE) and the mean average percentage error (MAPE) were also computed as secondary performance metrics. The same ADAM optimizer as used in the previous classification task was employed.

The images were resized to 256 pixels high and 512 pixels wide, and the pixel value was normalized to between 0 and 1. The same data augmentation procedure used in the classification task was also applied in this task. Batch size, number of epochs, termination condition and conditions for saving the final model state were also the same as those used in the classification task.

Operation advice system

A total of 1,070 concatenated images were contained in the dataset for suggesting the operation plan. In total, 750 (70%) belonged to the training set, 160 (15%) belonged to the validation set and 160 (15%) belonged to the retrospective testing set.

The same CNN architecture, loss function and secondary performance metrics as those used in the deviation evaluation task were employed. The images were resized to 256 pixels high by 1,024 pixels wide. The same data augmentation was applied to the training set. Batch size, number of epochs, termination condition and conditions for saving the final model state were the same as those used previously.

Model testing

In the first stage, two independent test sets were used to retrospective assess the performance of the CNN model: 780 photos for the screening system and 249 photos for the deviation evaluation system.

In the first stage, three independent test sets (as mentioned in the "Development of the algorithm" section) were used to retrospective assess the performance of

© Annals of Translational Medicine. All rights reserved.

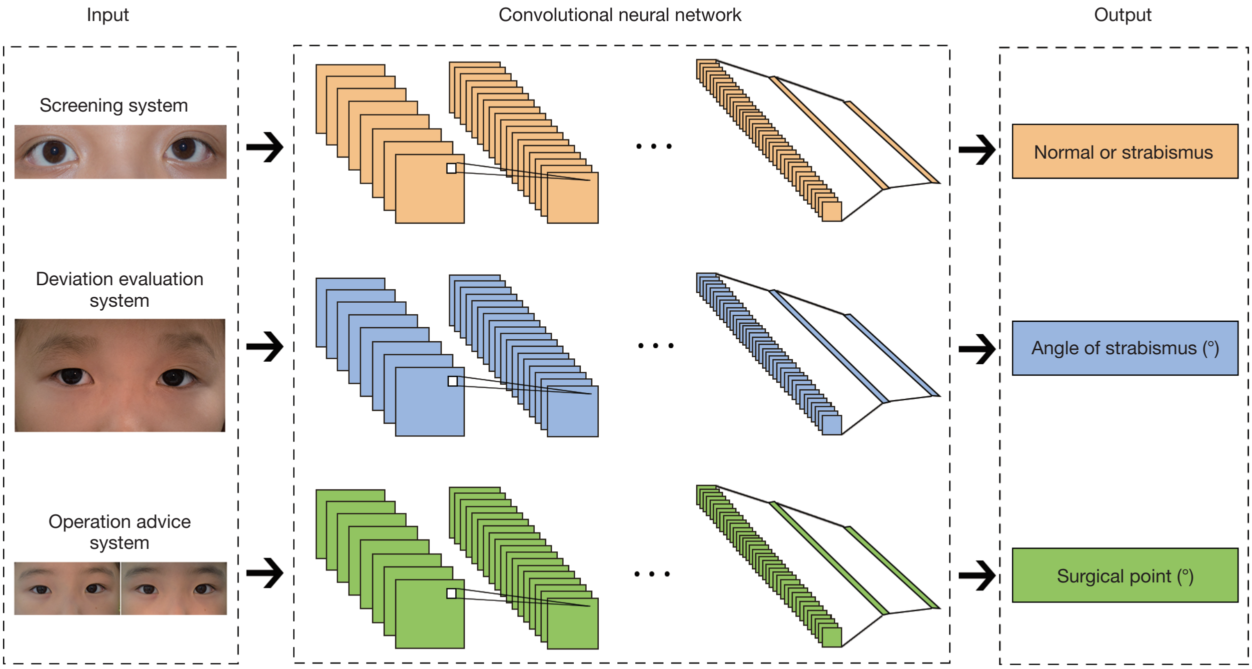


Figure 3 Diagram showing the framework of the AI platform. Three independent DL systems were established to screen for strabismus, evaluate deviation and propose an operation plan based on corneal light-reflection photos. AI, artificial intelligence; DL, deep learning.

the CNN model: 780 photos for the screening system, 249 photos for the deviation evaluation system, and 150 concatenated images for the operation evaluation system. To test the AI agent's screening ability when facing a realistic test, a "finding a needle in a haystack" test (27), which was composed of three independent groups (each contains 50 normal cases and a case with strabismus) derived from the previously described school, was include.

To further validate the clinical implementation of our AI platform, these three systems were prospectively tested in the second stage with populations encompassing various images.

Statistical analysis

Statistical analyses were performed using SPSS for Windows (Ver. 22.0, Statistical Package for the Social Sciences; SPSS, Inc., Chicago, USA) and the MedCalc statistical packages (Ver. 11.3, MedCalc Statistical Software, Ostend, Belgium). The performance of the screening system was evaluated in terms of area under the receiver operating characteristic curve (AUC) with 95% confidence intervals (95% CI). Our evaluation metrics also include sensitivity, specificity, and accuracy: where "sensitivity" means the ability of the algorithm to correctly identify patients with strabismus, sensitivity = true positive (TP)/[TP + false negative (FN)]; "specificity" means the ability to correctly identify people without strabismus, specificity = true negative (TN)/[TN + false positive (FP)]; and "accuracy" means the ability to correctly diagnose, accuracy = (TP + TN)/(TP + TN + FP + FN). The consistent variability between the evaluation system and the perimeter arc was represented using a Bland-Altman plot, while the correlation between them was calculated using Pearson and Spearman correlation coefficients. The same method was applied to analyze the agreement between the operation advice system and the actual designation. P values of <0.05 were considered statistically significant.

The main data supporting the results in this study are available within the paper. The raw and analysed datasets generated during the study are too large to be publicly shared, yet they are available for research purposes from the corresponding authors on reasonable request. The code is available from the corresponding authors on reasonable request as well.

Results

A total of 4,704 corneal light-reflection photos obtained from 2,154 patients were included in the hospital-based

© Annals of Translational Medicine. All rights reserved.

dataset, and 1,128 corneal light-reflection photos obtained from 1,125 individuals were included in the population-based dataset in the first stage. In the hospital-based dataset, 4,660 images were obtained preoperatively from 2,128 (98.79%) patients as a routine procedure prior to the operation in our clinics, and 44 images were obtained from 44 (2.04%) patients postoperatively during follow-up in our outpatient clinics. In the second stage, a total of 571 corneal light-reflection photos obtained from 323 individuals were included, and the overall subject demographics and image characteristics of the training, validation, and test datasets are listed in *Table 1*.

Retrospective testing

The retrospective test set for the screening system comprised 698 photos obtained from 356 patients with different types of strabismus (30.3% with esotropia, 67.7% with exotropia, and 2.0% with other types) and 172 photos from individuals with alignment; the mean age of the subjects was 13.4±7.9 years, and the sexes of the subjects were balanced. The screening system possesses the capacity to serve as an effective screening tool for the general population, achieving a sensitivity of 99.1% (95% CI, 98.0–99.7%), a specificity of 98.3% (95% CI, 94.6–99.5%), and an accuracy of 99.0% (95% CI, 98.0–99.5%). The AUC of the DL algorithm was 0.998 (95% CI, 0.993–1.000, P<0.001, *Figure 4*). In the "finding a needle in a haystack" test, the agent successfully identified the strabismus case in three rounds (exotropia in group 1, esotropia in group 2, and vertical strabismus in group 3, *Figure 5*).

Among the 249 photos included for retrospective testing of deviation evaluation system, approximately one-third were obtained from patients with esotropia and a mean angle of deviation of 18.9° (range, 4–40°) while two-thirds were obtained from patients with exotropia and a mean angle of deviation of 20.1° (range, 5–38°). The mean age of the 224 patients enrolled in the test set was 12.0±8.8 years and nearly half of them (49.1%) were female. The system and perimeter arc showed an excellent positive correlation (r=0.95, P<0.001, *Figure 6A*). The estimated angle of deviation was measured to within ±6.6° versus the perimeter arc, which was defined as the 95% limits of agreement (95% LoA, *Figure 6B*). The overall average error for horizontal deviation was 2.9° with respect to the measurements obtained with the perimeter arc. When the degrees were converted into PD using the following equation: PD = tan (deviation, degrees) × 100, the error was 5.7 PD. Moreover, the overall average errors were 3.0° (6.0 PD) and 2.6° (5.1 PD) for outward (n=165) and inward (n=84) deviations, respectively. For 214 (85.9%) of the 249 subjects, the difference in ocular deviation was <5° between the evaluation system and the perimeter arc.

Of the 160 subjects included in the retrospective test set for the operation advice system, 68.1% exhibited intermittent exotropia, 22.5% exhibited concomitant exotropia and 9.4% exhibited other relatively rare types of exotropia. The mean angle of exodeviation was 20.8° (range, 9–38°). The actual target angle and the surgical advice showed a high positive correlation (r=0.86, P<0.001, *Figure 6C*). The average difference between the prediction and the real target angle was 2.3° (4.7 PD). Compared to the target angle, this system achieved a level of accuracy of ±5.5° (11.5 PD) with a small bias of −0.6° (*Figure 6D*). Moreover, in the largest group (73.3%, n=110) in which the outward deviation ranged from 15.0° to 24.9°, the 95% LoA was ±4.5° (8.8 PD), suggesting that the system might potentially help ophthalmologists with determining the surgical design when applying a criterion for a successful operation for a deviation less than 10 PD (27).

Prospective testing

In total, 571 photos (506 with strabismus, 65 orthotropic) obtained from 323 patients were included in the test set for prospective testing of the screening system, and the mean age of the subjects was 14.8±11.6 years. The screening system showed an AUC of 0.980 (95% CI, 0.963–0.997, P=0.009, *Figure 4B*) with a sensitivity of 98.6% (95% CI, 97.1–99.4%), a specificity of 85.7% (95% CI, 74.1–92.9%) and an accuracy of 97.2% (95% CI, 95.5–98.3%).

Of the 202 photos included in the prospective test set for the deviation evaluation system, approximately half were obtained from patients with exotropia, and the rest were obtained from patients with exotropia. The mean angle was 19.6° (range, 8–32°) of exodeviation and 19.7° (range, 4–35°) of esodeviation. An excellent correlation (r=0.98, P<0.001, *Figure 6E*) was observed between the angles measured with the perimeter arc and those assessed by the algorithm. The estimated angle of deviation was measured to within ±7.0° (*Figure 6F*) versus the perimeter arc, and the overall average error was 2.6° (5.2 PD).

The prospective test set for the operation advice system comprised 38 subjects with intermittent exotropia, 11 with concomitant exotropia, and 7 with other relatively rare types of exotropia. The suggestion and the actual target

© Annals of Translational Medicine. All rights reserved.

Table 1 Overall subject demographics and image characteristics of the training, validation and prospective test datasets

Characteristics	First stage			Prospective test set
	Training set	Validation set	Retrospective test set	
Screening system				
Subject demographics				
Age, mean ± SD [range], y	13.8±8.1 [1–65]	14.0±8.4 [2–55]	13.4±7.9 [1–52]	14.8±11.6 [2–67]
Female, No./total (%)	1,154/2,364 (48.8)	266/527 (50.5)	266/528 (50.4)	169/323 (52.3)
No. of strabismus	1,561	355	356	247
Esotropia, No, (%)	413 (26.5)	115 (32.4)	108 (30.3)	160 (61.8)
Exotropia, No. (%)	1,115 (71.4)	231 (65.1)	241 (67.7)	87 (33.6)
Other types, No. (%)	33 (2.1)	9 (2.5)	7 (2.0)	12 (4.6)
Photo characteristics				
With glasses, No./total (%)	444/4,057 (10.9)	94/870 (10.8)	94/870 (10.8)	60/571 (10.5)
Strabismus, No. (%)	3,254 (80.2)	698 (80.2)	698 (80.2)	506 (88.8)
Deviation evaluation system				
Subject demographics				
Age, mean ± SD [range], y	11.2±8.5 [1–55]	11.6±7.6 [4–49]	12.0±8.8 [3–48]	12.5±10.5 [2–53]
Female, No./total (%)	357/736 (48.5)	92/222 (41.4)	110/224 (49.1)	49/95 (51.6%)
Esotropia, No, (%)	227 (30.8)	70 (31.5)	74 (33.0)	35 (36.8)
Photo characteristics				
With glasses, No./total (%)	193/1165 (16.8)	47/249 (18.9)	48/249 (19.3)	34/202 (16.8)
Esotropia, No. (%)	419 (36.0)	86 (34.5)	84 (33.7)	85 (42.1)
Operation advice system				
Subject demographics				
Age, mean ± SD [range], y	12.0±7.5 [1–47]	11.8±7.9 [2–42]	11.4±8.2 [3–49]	10.9±7.4 [3–40]
Female, No./total (%)	328/750 (43.7)	82/160 (51.2)	77/160 (48.1)	31/56 (55.3%)
Type of strabismus				
Intermittent, No. (%)	425 (56.7)	93 (58.1)	109 (68.1)	38 (67.9)
Concomitant, No. (%)	240 (32.0)	51 (31.9)	36 (22.5)	11 (19.6)
with V pattern, No. (%)	54 (7.2)	9 (5.6)	12 (7.5)	3 (5.4)
with A pattern, No. (%)	13 (1.7)	4 (2.5)	1 (0.6)	1 (1.8)
Infantile, No. (%)	10 (1.3)	3 (1.9)	2 (1.3)	0 (0)
Sensory, No. (%)	8 (1.1)	0 (0.0)	0 (0.0)	3 (5.4)
Photo characteristics				
With glasses, No./total (%)	31/750 (4.1)	8/160 (5.0)	6/160 (3.8)	1/56 (1.8)

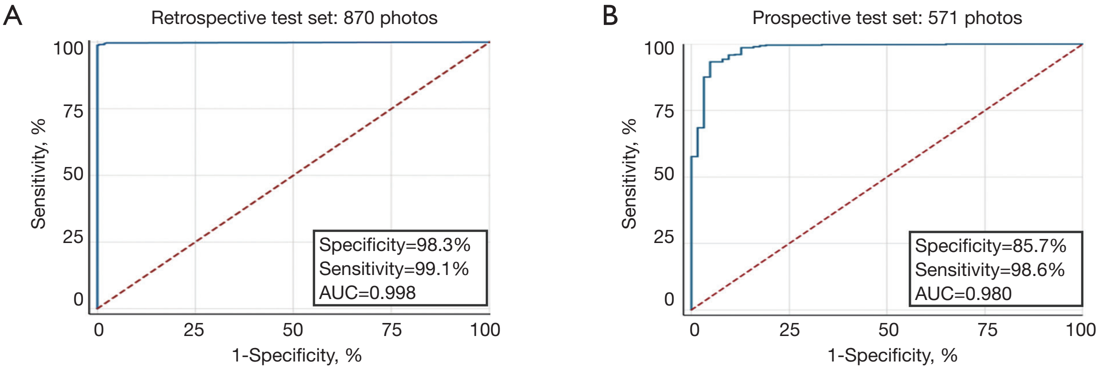


Figure 4 Performance of the screening system. (A) The receiver-operating characteristic (ROC) analysis graphically illustrates the excellent diagnostic performance of the algorithm (blue curve), with an area under the curve (AUC) of 0.998 against a random chance diagnosis (red solid line) derived from the retrospective test set. The sensitivity was 99.1%, and the specificity was 98.3% (n=870). (B) Graph showing an AUC of 0.980 that was obtained using the algorithm derived from the prospective test set. The sensitivity was 85.7%, and the specificity was 98.6% (n=571).

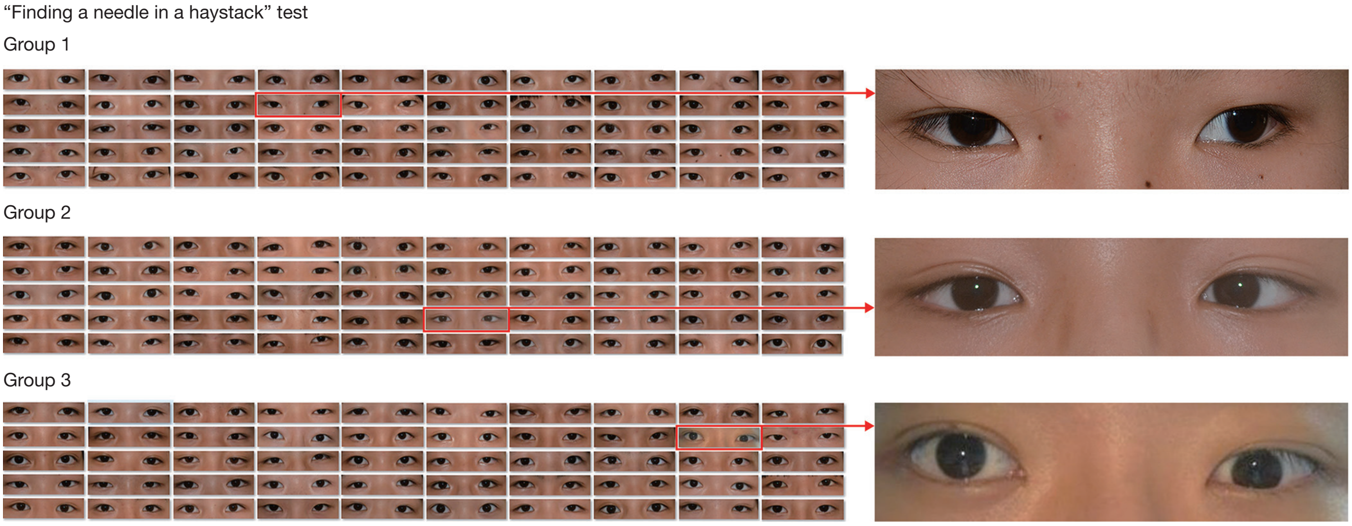


Figure 5 The screening system successfully identified the strabismus case in three rounds in the "finding a needle in a haystack" test. The right column of images shows the strabismus case in each group.

angle showed a good positive correlation (r=0.76, P<0.001, *Figure 6G*). Compared to the target angle, this system achieved a level of accuracy of ±6.1° (*Figure 6H*), and the average difference between the prediction and the real target angle was 2.5° (5.0 PD).

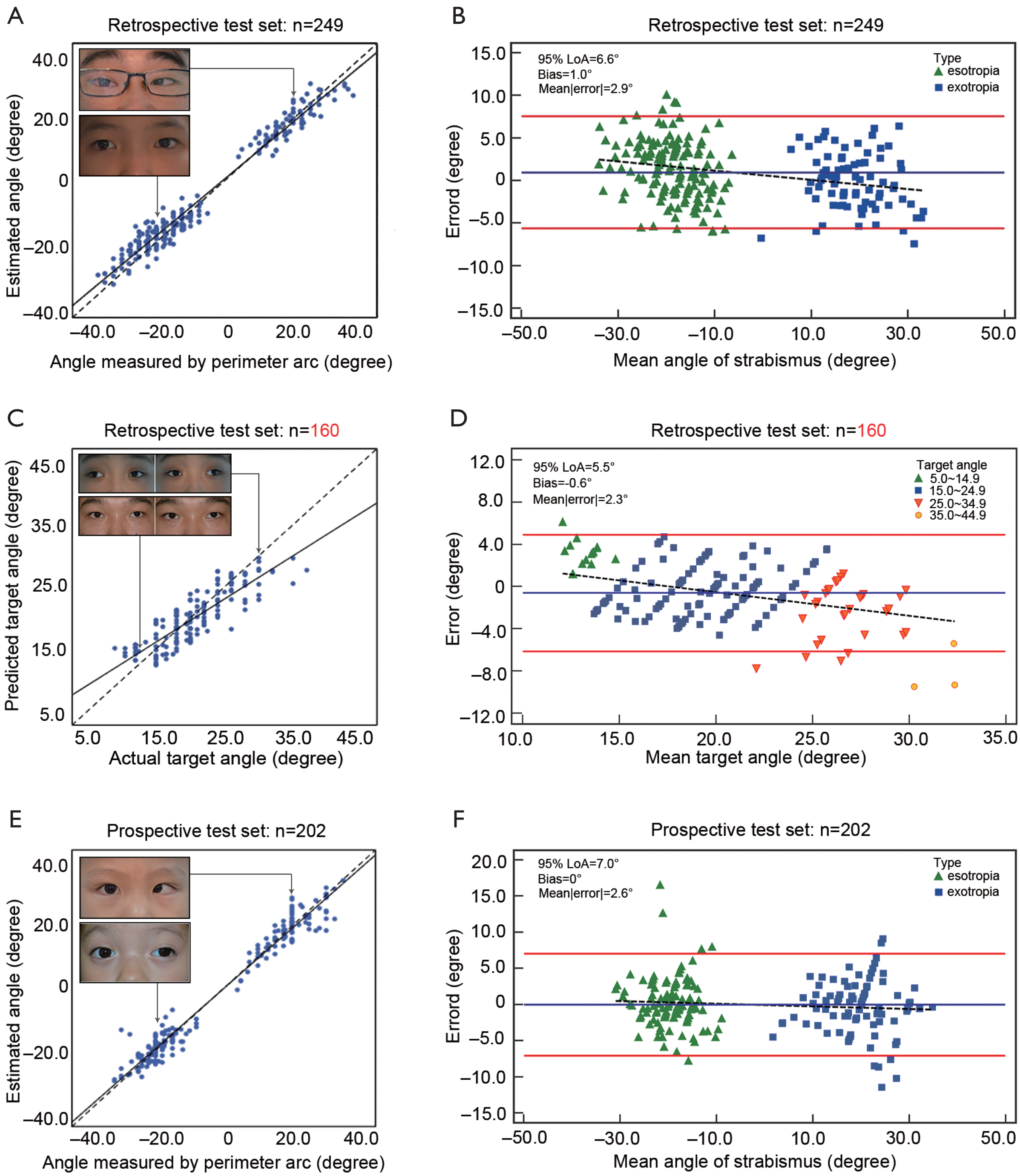
Discussion

In this study, we established an AI platform to automatically screen for strabismus, evaluate deviation and provide surgical advice based on corneal light-reflection photos. Our screening system achieved 99.1% sensitivity, 98.3% specificity, and 99.0% matching in the retrospective test, which were comparable with or better than the results obtained using previously established methods for the automatic detection of strabismus based on digital images (94.17% accuracy, 97.23% sensitivity, and 73.08% specificity) (28) or digital videos (100% specificity, 80% sensitivity, and 93.33% accuracy for exotropia) (20). Our algorithm also showed a robust performance (AUC, 0.980,

© Annals of Translational Medicine. All rights reserved.

sensitivity 98.6%, specificity 85.6%) in the retrospective test on 571 photos which were obtained with three different devices, indicating a potential application in different forms of images taken by different photographers.

Current strabismus evaluation technologies are mostly designed for horizontal strabismus, which accounts for 40.57–98% of affected cases (29,30) and are heavily dependent on model eyes. However, the constant Hirschberg ratio used by many 2D models to convert the distance from the reflex center to the limbus center (RD) into a strabismus angle is subject to high interindividual variations (±20% of the mean value), which may cause



© Annals of Translational Medicine. All rights reserved.

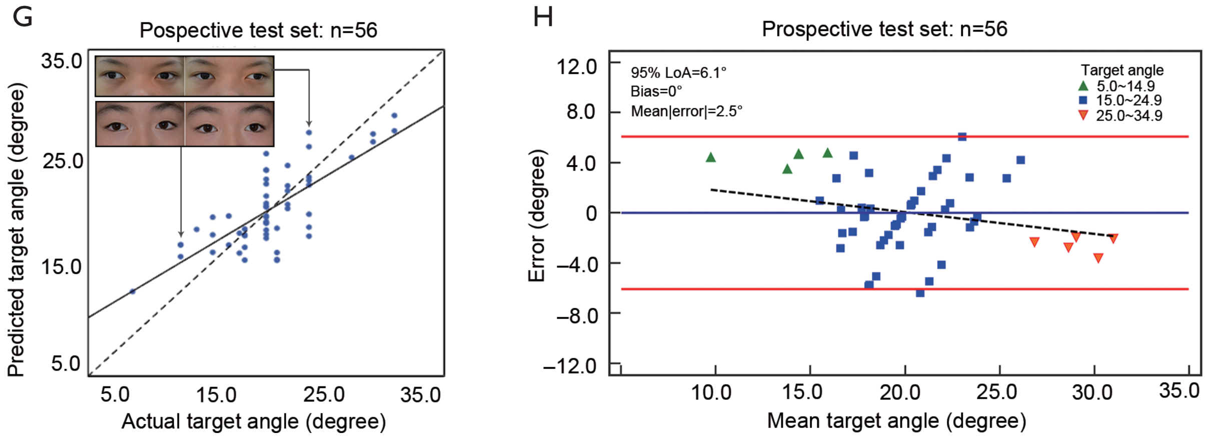


Figure 6 Performance of the deviation evaluation system and the operation advice system. (A,B,C,D) shows the performance based on the retrospective test set while (E,F) shows the results derived from the prospective test set. (A) The deviation evaluation system measured the horizontal strabismus angle within ±2.9° of the angle measured based on the perimeter arc (r=0.95, mean |error|). The solid line represents the ideal results while the dashed line represents the actual data fit. (B) The Bland-Altman analysis revealed a bias with an average error of 1.0°. The dashed line represents the relationship between the residual and the average strabismus angle measurements obtained from the perimeter arc and the algorithm (r=–0.05); the solid red lines represent the 95% limits of agreement (±6.6°). (C) The system provided advice regarding the target angle that was within ±2.3° of the actual target angle (r=0.86, mean |error|). (D) The Bland-Altman analysis revealed that compared to the actual target angle, the system achieves a level of accuracy of ±5.5° (95% LoA), with a small bias of –0.6°. (E) The deviation evaluation system measured the horizontal strabismus angle to within ±2.6° of the angle measured based on the perimeter arc (r=0.98, mean |error|). (F) The Bland-Altman analysis revealed that the deviation evaluation system achieves a level of accuracy of ±7.0° (95% LoA) compared to the angle measured with the perimeter arc. (G) The system provided advice regarding the target angle that was within ±2.5° of the actual target angle (r=0.76, mean |error|). (H) The Bland-Altman analysis revealed that compared to the actual target angle, the system achieves a level of accuracy of ±6.1° (95% LoA).

significant errors (31,32). For example, a system developed based on adult data may not be suitable for application in children. Even age-adjusted 3D eye models based on ophthalmic biometric data may not fully represent individual characteristics, limiting their applications in subjects with high refractive errors (17,33,34). In addition, errors in the localization and segmentation may occur during feature extraction, resulting in misclassification or an inconclusive result.

Instead of building a model eye with constant parameters, we applied a CNN, a class of DL networks, in our study to avoid the weaknesses of model eye designs. CNNs are biologically inspired variants of multilayer perceptrons and tend to recognize visual patterns directly from raw image pixels (22,35). Recently, Chen *et al.* (36) applied CNNs to strabismus screening using six different models and showed that the VGG-S model had the best specificity (96%) and sensitivity (94.1%). To train CNN models with high reliability, a larger eye-tracking dataset with a separate test set is required. To our knowledge, this is the first study to apply CNN in strabismus based on a single photo. The advantages of DL can be fully demonstrated in our study using big data, although interpreting the resulting model remains difficult (similar to a "black box"). A strength of this study is the enrollment of more than 3000 subjects in our dataset, and the population was approximately 20 times larger than the populations included in previous studies, thus ensuring the accurate detection of strabismus on a large scale (17-19,25,33,37).

Only one photograph is needed for our model in the screening and deviation evaluation. Corneal light-reflection photos, which are noninvasive and require minimal cooperation, are relatively easy to acquire even in young children. Valente *et al.* (20) measured the strabismus angle based on digital videos of cover tests with at least 5 iterations, and the average error was 2.57 PD compared with the PCT, while the VOG test requires video goggles to be worn for approximately 2 min and has a 95% LoA of ±5.05 PD (19). Images acquired with a special infrared camera were required in another study that achieved a 95%

© Annals of Translational Medicine. All rights reserved.

Table 2 Surgical dose of rectus resection or recession for patients with exotropia

Operations	Target angle (º)	LR recession (mm)	MR resection (mm)
Bilateral lateral rectus recession	15	5+5	NA
	16.5	5+6	NA
	18	6+6	NA
	21	7+7	NA
Bilateral medial rectus resection	35	NA	7+7
	40	NA	8+8
Unilateral lateral recession with medial rectus resection	7.5	5	0
	9	6	0
	10	0	4
	12	8	0
	13.5	9	0
	17.5	0	7
	18	5	4
	20	5	5
	22	6	5
	24	6	6
	26	7	6
	28	7	7
	30	8	7
	32	8	8
	34	9	8
Bilateral lateral recession with unilateral medial rectus resection	36.5	7+7	6
	38	8+8	5
	38.5	7+7	7
	42	8+8	7

LoA of ±8.5 PD (34). However, the long test time and/or high expenses associated with the equipment may limit their application in real clinics.

In our practice, 1,000 photos of participants were obtained within a 2-hour time frame by a single photographer, indicating the potential for high-volume analysis compared with video acquisition in other studies, which requires more than 15 s to capture per participant (19,20). Errors were smaller than 10 PD in 212 (85.14%) patients with an overall average of 2.9° (5.7 PD) in the retrospective test set and in 180 (89.1%) patients with an overall average of 2.6 (5.2 PD) in the prospective test set. Variations less than 10 PD in APCT measurement, which is the gold standard for angle measurement, are likely due to interexaminer variability (12,38); therefore, our platform can provide credible references for misalignment assessment.

Extraocular muscle surgery design always involves two steps. The first step is the determination of the target angle, which is defined as the deviation we plan to eliminate. The target angle is derived from the measurement of deviation in different circumstances (e.g., deviation with either eye staring forward). The second step is the choice of the surgical technique (e.g., bilateral lateral rectus muscle recessions or unilateral recession and resection) (8,39,40) and surgical dose (39,41,42), which may vary according to the preoperative diagnosis, the angle of deviation at distance and near, the anatomy of the extraocular muscle, physician preference and experience (8). The procedure was undertaken using the surgical dose at the authors' clinic as shown in *Table 2*. As surgical designs can be confusing for ophthalmologists despite the debate regarding surgical techniques (6), we aimed to integrate concatenated corneal light-reflection photos and the experience of experts (the actual target angle in each operation) by training an algorithm to provide advice regarding the target angle. In practice, the angle of deviation at distance and near are also considered in addition to the deviation recorded with either eye staring forward when determining the target angle (6). Considering the limited information about near distance provided by the photos, our model cannot provide a definitive target angle but rather serves as a reference for surgical planning. Our model is currently suitable for constant, intermittent, perceptual exotropia and exotropia with a "V" or "A" pattern; subjects with secondary and paralytic strabismus will still require personalized designs by experts. The average difference between the forecast and actual target angles in the retrospective test set was 2.3° (4.7) in all subjects and 2.0° (3.9 PD) in the most popular group (angle=15.0–24.9°), which is smaller than the angle interval of 5 PD for surgical procedures used in most protocols (41,42). As exotropia is the most common subtype of strabismus in China and the primary type found in inpatients who need surgery (4), the operation advice system in our platform may provide substantial help in clinical practice.

Some limitations of our study must be considered. First,

© Annals of Translational Medicine. All rights reserved.

our photos were obtained in an Asian population. A larger dataset covering a wider variety of ethnic groups would further improve our system and validate its performance in a broader ethnic population. Second, our deviation evaluation system is available for horizontal strabismus, while the operation advice system is currently suitable for a majority of types of exotropia. Further studies are needed to extend the application ranges and enroll more parameters such as information on far distance to improve the operation advice system. Given the performance of our AI platform in the prospective test, we plan to release an application for mobile phones to automatically screen strabismus in a more convenient manner and further assess the screening system in our future studies. And our work is best understood in the context that the photos used were taken by photographers, and when it comes to clinical implication, the quality of the photos must be taken into consideration. Like other deep neuron networks, our platform is also prone to mistakes that humans are much less likely to make. Part of them may be due to insufficient information given by a single photo, while others are unexplainable. After all. we always believe that cooperation between human specialists and AI can achieve better performance than either individually.

Conclusions

In conclusion, our AI platform achieved excellent performance with high accuracy and requires only corneal light-reflection photos, which are easily obtainable using a digital camera. Given the high global prevalence of strabismus, our system has significant implications as a widely accessible screening tool for the general population. In addition, our deviation evaluation system and operation advice system will substantially benefit doctors in general hospitals, especially in underdeveloped areas with limited resources.

Acknowledgments

Funding: This study was funded by the National Key Research and Development Program of China (2018YFC0116500), the National Natural Science Foundation of China (81770967), the Science and Technology Planning Projects of Guangdong Province (2017B030314025, 2018B010109008, 2019B030316012) and Guangdong Science and Technology Innovation Leading Talents (2017TX04R031). The sponsor or funding organization had no role in the design or conduct of this research.

Footnote

Reporting Checklist: The authors have completed the Materials Design Analysis Reporting (MDAR) checklist and the STROBE reporting checklist. Available at http://dx.doi.org/10.21037/atm-20-5442

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-20-5442

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-20-5442). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was approved by the Institutional Review Board of Zhongshan Ophthalmic Center (2019KYPJ153) and adhered to the tenets of the Declaration of Helsinki (as revised in 2013). Written informed consent was obtained from the participants or their legal guardians

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Prevalence of amblyopia and strabismus in African American and Hispanic children ages 6 to 72 months the multi-ethnic pediatric eye disease study. Ophthalmology 2008;115:1229-36.e1.
2. Frandsen AD. Occurrence of squint: a clinical-statistical study on the prevalence of squint and associated signs in different groups and ages of the Danish population. Acta Ophthalmol 1960;62:1.
3. Graham PA. Epidemiology of strabismus. Br J Ophthalmol 1974;58:224-31.
4. Chen X, Fu Z, Yu J, et al. Prevalence of amblyopia and

strabismus in Eastern China: results from screening of preschool children aged 36-72 months. Br J Ophthalmol 2016;100:515-9.

5. McKean-Cowdin R, Cotter SA, Tarczy-Hornoch K, et al. Prevalence of amblyopia or strabismus in asian and non-Hispanic white preschool children: multi-ethnic pediatric eye disease study. Ophthalmology 2013;120:2117-24.
6. Wallace DK, Christiansen SP, Sprunger DT, et al. Esotropia and Exotropia Preferred Practice Pattern(R). Ophthalmology 2018;125:P143-83.
7. Birch EE, Fawcett S, Stager DR. Why does early surgical alignment improve stereoacuity outcomes in infantile esotropia? J AAPOS 2000;4:10-4.
8. Oliva O, Morgado A. Bilateral lateral rectus recession versus unilateral recession/resection for basic intermittent exotropia. Medwave 2018;18:e7319.
9. Sharma P, Gaur N, Phuljhele S, et al. What's new for us in strabismus? Indian J Ophthalmol 2017;65:184-90.
10. Read JC. Stereo vision and strabismus. Eye (Lond) 2015;29:214-24.
11. Wright KW, Spiegel PH. Pediatric ophthalmology and strabismus, second ed. Springer Science and Business Media, 2013.
12. de Jongh E, Leach C, Tjon-Fo-Sang MJ, et al. Inter-examiner variability and agreement of the alternate prism cover test (APCT) measurements of strabismus performed by 4 examiners. Strabismus 2014;22:158-66.
13. Du D, Xiongwu Z, Zhenhua D, et al. Analysis of diversity between perimeter strabismometry and triple prism strabismometry. Chinese Journal of Strabismus & Pediatric Ophthalmology 2011;19:64-7.
14. Liu L, L. H, F. C, et al. Design of the digital arc perimeter applied in concomitant strabismus. Chinese Journal of Strabismus & Pediatric Ophthalmology 2016;24:262-5.
15. Hu YH, Jia HL. Application of prism alternate cover test and perimeter strabismometry in strabismic operation. Guoji Yanke Zazhi (Int Eye Sci) 2015;15:1290-2.
16. Choi RY, Kushner BJ. The accuracy of experienced strabismologists using the Hirschberg and Krimsky tests. Ophthalmology 1998;105:1301-6.
17. Yang HK, Seo JM, Hwang JM, et al. Automated analysis of binocular alignment using an infrared camera and selective wavelength filter. Invest Ophthalmol Vis Sci 2013;54:2733-7.
18. Sousa de Almeida JD, Silva AC, Teixeira JA, et al. Computer-Aided Methodology for Syndromic Strabismus Diagnosis. J Digit Imaging 2015;28:462-73.
19. Park N, Park B, Oh M, et al. A quantitative analysis method for comitant exotropia using video-oculography with alternate cover. BMC Ophthalmol 2018;18:80.
20. Valente TLA, de Almeida JDS, Silva AC, et al. Automatic diagnosis of strabismus in digital videos through cover test. Comput Methods Programs Biomed 2017;140:295-305.
21. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, et al. Artificial intelligence in retina. Prog Retin Eye Res 2018;67:1-29.
22. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019;103:167-75.
23. Li Z, Guo C, Nie D, et al. Deep learning for detecting retinal detachment and discerning macular status using ultra-widefield fundus images. Commun Biol 2020;3:15.
24. He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30-6.
25. Almeida JD, Silva AC, Paiva AC, et al. Computational methodology for automatic detection of strabismus in digital images through Hirschberg test. Comput Biol Med 2012;42:135-46.
26. Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:1602.07261.
27. Long E, Lin H, Liu Z, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nature Biomedical Engineering 2017;1:0024.
28. Seo MW, Yang HK, Hwang JM, et al. editors. The Automated Diagnosis of Strabismus Using an Infrared Camera. 6th European Conference of the International Federation for Medical and Biological Engineering; 2015; Cham: Springer International Publishing.
29. Bruce A, Santorelli G. Prevalence and Risk Factors of Strabismus in a UK Multi-ethnic Birth Cohort. Strabismus 2016;24:153-60.
30. Repka MX, Lum F, Burugapalli B. Strabismus, Strabismus Surgery, and Reoperation Rate in the United States: Analysis from the IRIS Registry. Ophthalmology 2018;125:1646-53.
31. Schaeffel F. Kappa and Hirschberg ratio measured with an automated video gaze tracker. Optom Vis Sci 2002;79:329-34.
32. Model D, Eizenman M, Sturm V. Fixation-free assessment of the Hirschberg ratio. Invest Ophthalmol Vis Sci 2010;51:4035-9.
33. Yang HK, Han SB, Hwang JM, et al. Assessment of binocular alignment using the three-dimensional Strabismus

© Annals of Translational Medicine. All rights reserved.

Photo Analyzer. Br J Ophthalmol 2012;96:78-82.

34. Yoo YJ, Yang HK, Seo JM, et al. Infrared photographs with a selective wavelength filter to diagnose small-angle esotropia in young children. Graefes Arch Clin Exp Ophthalmol 2019;257:645-50.
35. Anwar SM, Majid M, Qayyum A, et al. Medical Image Analysis using Convolutional Neural Networks: A Review. J Med Syst 2018;42:226.
36. Chen Z, Fu H, Lo WL, et al. Strabismus Recognition Using Eye-Tracking Data and Convolutional Neural Networks. J Healthc Eng 2018;2018:7692198.
37. Dericioğlu V, Cerman E. Quantitative measurement of horizontal strabismus with digital photography. J AAPOS 2019;23:18.e1-6.
38. Interobserver reliability of the prism and alternate cover test in children with esotropia. Arch Ophthalmol 2009;127:59-65.
39. Pediatric Eye Disease Investigator Group; Writing Committee, Donahue SP, et al. A Randomized Trial Comparing Bilateral Lateral Rectus Recession versus Unilateral Recess and Resect for Basic-Type Intermittent Exotropia. Ophthalmology 2019;126:305-17.
40. Xie F, Zhao K, Zhang W. Comparison of surgical outcomes between bilateral recession and unilateral recession-resection in moderate-angle intermittent exotropia. J AAPOS 2019;23:79.e1-79.e7.
41. Jung EH, Yu YS, Kim SJ. A comparison of surgical outcomes between pre-and full-term patients with exotropia. PLoS One 2018;13:e0208848.
42. Kim WJ, Kim MM. The fast exodrift after the first surgical treatment of exotropia and its correlation with surgical outcome of second surgery. BMC Ophthalmol 2018;18:67.

Cite this article as: Mao K, Yang Y, Guo C, Zhu Y, Chen C, Chen J, Liu L, Chen L, Mo Z, Lin B, Zhang X, Li S, Lin X, Lin H. An artificial intelligence platform for the diagnosis and surgical planning of strabismus using corneal light-reflection photos. Ann Transl Med 2021;9(5):374. doi: 10.21037/atm-20-5442

© Annals of Translational Medicine. All rights reserved.