tvst

Real-Time Risk Score for Glaucoma Mass Screening by Spectral Domain Optical Coherence Tomography: Development and Validation

Kota Fukai[1,*], Ryo Terauchi[2,*], Takahiko Noro[2], Shumpei Ogawa[2], Tomoyuki Watanabe[2], Toru Nakagawa[3], Toru Honda[3], Yuya Watanabe[3], Yuko Furuya[1], Takeshi Hayashi[3], Masayuki Tatemichi[1], and Tadashi Nakano[2]

[1] Department of Preventive Medicine, Tokai University School of Medicine, Kanagawa, Japan
[2] Department of Ophthalmology, The Jikei University School of Medicine, Tokyo, Japan
[3] Hitachi Health Care Center, Ibaraki, Japan

Correspondence: Tadashi Nakano, Department of Ophthalmology, The Jikei University School of Medicine, 3-25-8 Nishi-Shimbashi, Minato-ku, Tokyo 105-8461, Japan. e-mail: tnakano@jikei.ac.jp

Purpose: To develop and validate a risk score assessable in real-time using only retinal thickness-related values measured by spectral domain optical coherence tomography alone for use in population-based glaucoma mass screenings.

Methods: A total of 7572 participants (aged 35–74 years) underwent spectral domain optical coherence tomography examination annually between 2016 to 2021 in a population-based setting. We selected 284 glaucoma cases and 284 controls, matched by age and sex, from 11,487 scans in 2016. We conducted multivariable logistic regression with backward stepwise selection of retinal thickness-related variables to develop the diagnostic models. The developed risk scores were applied to all participants in 2018 (9720 eyes), and we randomly selected 723 scans for validation. Additional validation using the Humphrey field analyzer was conducted on 129 eyes in 2020. We assessed the models using sensitivity, specificity, the area under the receiver operating characteristic curve and positive and negative predictive values.

Results: The best-predicting model achieved an area under the receiver operating characteristic curve of 0.97 (95% confidence interval, 0.96–0.98) with a sensitivity of 0.93 and specificity of 0.91. The validation dataset showed a positive predictive value of 90.8% for high-risk scorers, corresponding to 6.2% of the population, and negative predictive value of 88.2% for low-risk scorers, corresponding to 85.2%. Sensitivity and specificity for glaucoma diagnosis were 0.85 and 0.91, when we set the risk score cut-off at 90 points out of 100.

Conclusions: This risk score could be used as a valid index for glaucoma screening in a population-based setting.

Translational Relevance: The score is feasible by installing a simple computer application on an existing spectral domain optical coherence tomography and will help to improve the accuracy and efficiency of glaucoma screening.

Introduction

Glaucoma remains the major leading cause of blindness worldwide.[1] The worldwide prevalence of glaucoma, including primary open-angle glaucoma and primary angle-closure glaucoma, was estimated to be 3.54% in 2013, for a total of 64.3 million cases among those aged 40 to 80 years, and this number is expected to increase to 111.8 million in 2040.[2] Despite its high prevalence, previous epidemiological studies demonstrated that 50% to 90% of people with glaucoma are undiagnosed and are unaware that they have glaucoma.[3,4]

To prevent restrictions on daily life from visual field loss and blindness, it is important to detect the

disease at an early stage so that it can be treated in an early and appropriate manner.[5] Glaucoma mass screening in population-based settings using single or multiple ophthalmologic examinations is expected to prevent reduction in vision related quality of life from glaucoma progression.[6–9] Although glaucoma mass screening programs are not common in most developed countries, two recent studies from China[10] and India,[11] home to 40% of the world's population with glaucoma,[12] have found that opportunistic glaucoma mass screening in their context was cost effective. Although cost effectiveness has been shown in these screening programs, there still have issues with reliability in identifying glaucoma in population-based settings.[13] Another issue of normal-tension glaucoma, showing normal range with intraocular pressure, is present among several races, particularly in Asia.[3,14] Normal-tension glaucoma cannot be screened by measurement of intraocular pressure. Therefore, a feasible and reliable perimetric test or fundus examination suitable for glaucoma mass screening is required.[15,16]

Optical coherence tomography (OCT) is a noncontact, fast, and noninvasive technique that allows high-resolution cross-sectional imaging of the optic nerve head and retina.[17–19] OCT could be used by both general physicians in addition to ophthalmologists or as a screening tool in a community setting.[20,21] More recently, spectral domain (SD)-OCT has been developed. This system completes scanning of an eye within 1 minute without the need for specialist technicians.[22] Several groups, including ours, have sought algorithms that allow SD-OCT to be used for the mass screening for glaucoma. Among the case-control studies conducted to date, these groups achieved relatively high performance and clear separation between normal and glaucoma cases.[22–24] This success has led, in turn to the development and implementation of artificial intelligence–based algorithms for SD-OCT in clinical-based settings.[25–27] However, when applied to mass screening at population-based levels, these algorithms have been strongly handicapped by high rates of false positives.[24,26,28] No SD-OCT based screening algorithm is currently in mass screening use at a population-based level.

The purpose of this study was to develop and validate an algorithm for population-based glaucoma mass screening that can be applied to an existing SD-OCT. First, we verified the logic in detail of how ophthalmologists diagnose the OCT images as glaucomatous. Next, according to the logic, considerable patterns of variables in SD-OCT data were selected and we attempted to improve accuracy using regression analysis to develop glaucoma risk scores. Then, the glaucoma risk score was validated across population-based settings using two separate datasets of participants diagnosed by ophthalmologists, based only on OCT reports and on both OCT reports and Humphrey Field Analyzer (HFA) (Carl Zeiss AG, Oberkochen, Germany). To make an accurate diagnosis, because glaucoma is a progressive disease, we thought that it was better to consider the passage of time.

Methods

Informed consent was obtained from all participants by electronic form. The study adhered to the tenets of the Declaration of Helsinki.[29] This diagnostic study followed the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) reporting guideline.[30] The study protocol was approved by the ethics review boards of Hitachi Hospital (2014-63) and Tokai University School of Medicine (19R-090).

Study Design and Population

Details of the study setting are described elsewhere.[22,31] The flow chart of the development and validation process in this study is described in Figure 1. The development data was derived from participants who underwent the annual health checkup in fiscal year (FY) 2016. All participants in the development data (11,487 eyes of 7572 participants; mean age 51.3 ± 10.0 years; range, 35–74 years) underwent a comprehensive ophthalmologic examination which consisted of a review of medical history, digital fundus photography (Maestro, Topcon Corp., Tokyo, Japan), frequency doubling technology perimetry tests (screening mode C-20-1; Carl Zeiss AG, Oberkochen, Germany), and automatic OCT measurements (3D OCT-1 Maestro, Topcon Corp.). Axial length was measured using an optical axial length meter (Aladdin, Topcon Corp.). For the development of glaucoma risk scores, we used a matched case-control design. Cases were those diagnosed by ophthalmologists based on complete perimetric and fundus examinations and whose OCT reports were confirmed to be glaucomatous by ophthalmologists (RT and TaNa). Eligible controls were those who answered that had no history of glaucoma and whose fundus photography and frequency doubling technology tests showed no findings of glaucoma. We randomly selected one control for each case by matching with sex (male or female) and age (5-year strata). The final development dataset used for analysis
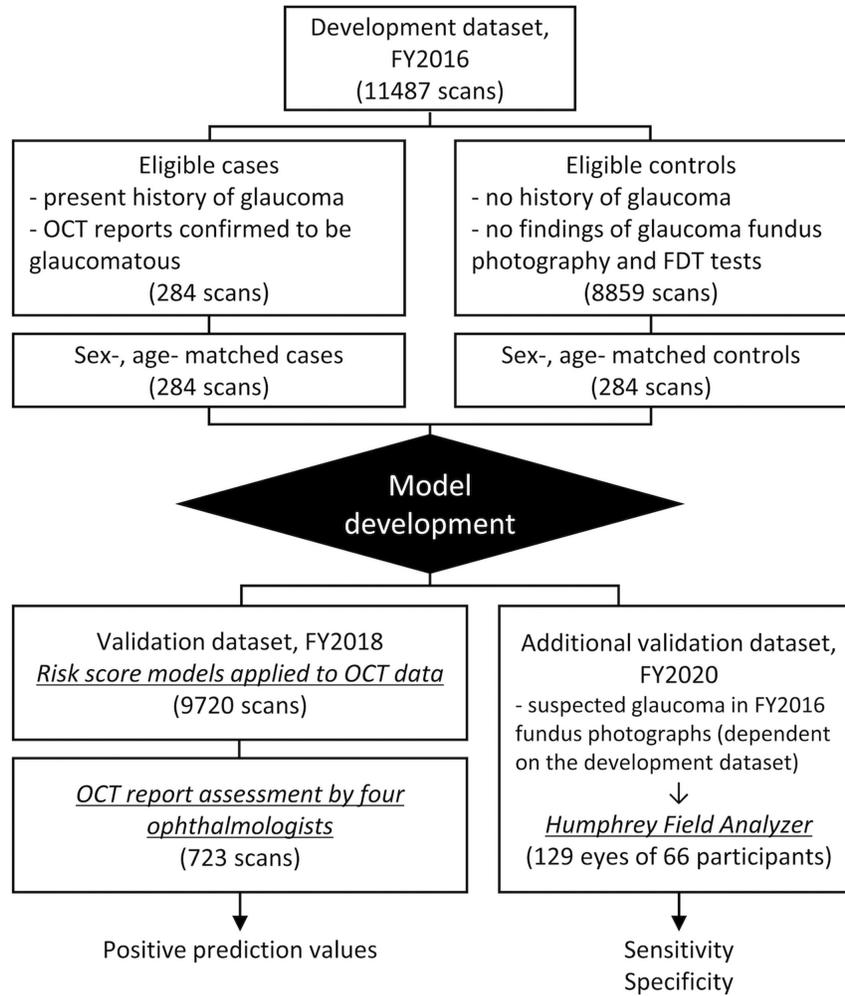
```
                        Development dataset,
                            FY2016
                        (11487 scans)

    Eligible cases                          Eligible controls
    - present history of glaucoma           - no history of glaucoma
    - OCT reports confirmed to be           - no findings of glaucoma fundus
    glaucomatous                            photography and FDT tests
    (284 scans)                             (8859 scans)

    Sex-, age- matched cases                Sex-, age- matched controls
    (284 scans)                             (284 scans)

                        ◆ Model development ◆

    Validation dataset, FY2018              Additional validation dataset,
    Risk score models applied to OCT data       FY2020
    (9720 scans)                            - suspected glaucoma in FY2016
                                            fundus photographs (dependent
                                            on the development dataset)
                                                       ↓
    OCT report assessment by four           Humphrey Field Analyzer
    ophthalmologists                        (129 eyes of 66 participants)
    (723 scans)

    Positive prediction values              Sensitivity
                                            Specificity
```

Figure 1. Flow chart of the development and validation process. FDT, frequency doubling technology; OCT, optical coherence tomography.

consisted of 284 eyes of 191 participants for cases and 284 eyes of 277 participants for controls.

The validation data source was derived from participants who underwent an annual health checkup in FY2018, 2 years after the development step. Based on the development step, the model obtained by the statistical analysis (described elsewhere in this article) was applied to all OCT measurements of the participants in FY2018. For validation, we randomly selected 723 eyes in each score level (low, middle, high). We then assigned four ophthalmologists (RT, SO, TNo, and TW, supervised by TaNa) to independently evaluate the OCT reports and judge whether the reports were normal, glaucoma, eye diseases other than glaucoma, or could not be judged (requirement of further full ophthalmological examinations). The four ophthalmologists used the following OCT parameters to determine glaucomatous changes: (1) average of retinal nerve fiber layer (RNFL) circular thickness; (2) presence of focal RNFL thinning and difference in height between the

double humps in temporal–superior–nasal–inferior–temporal (TSNIT) plot; (3) the quadrant and clock hour RNFL thickness charts; (4) retina, ganglion cell layer (GCL)++, and GCL+ thickness map in the macular area and RNFL thickness map in optic disc area; and (5) GCL++, GCL+, and RNFL thickness deviation map. Glaucomatous RNFL loss was often seen as an arcuate-type defect with temporal raphe. A decision meeting was held for subjects whose evaluation did not agree, and the assessment was repeated until all four ophthalmologists agreed. All ophthalmologists who participated in the assessment of this validation data were blinded to the scores and had no access to any other values nor images other than the OCT report.

As an additional subset analysis for validation, we conducted a follow-up survey in FY2020 for those with suspected glaucoma in FY2016 fundus photographs (dependent on the development step). One hundred twenty-nine eyes of 66 participants were additionally
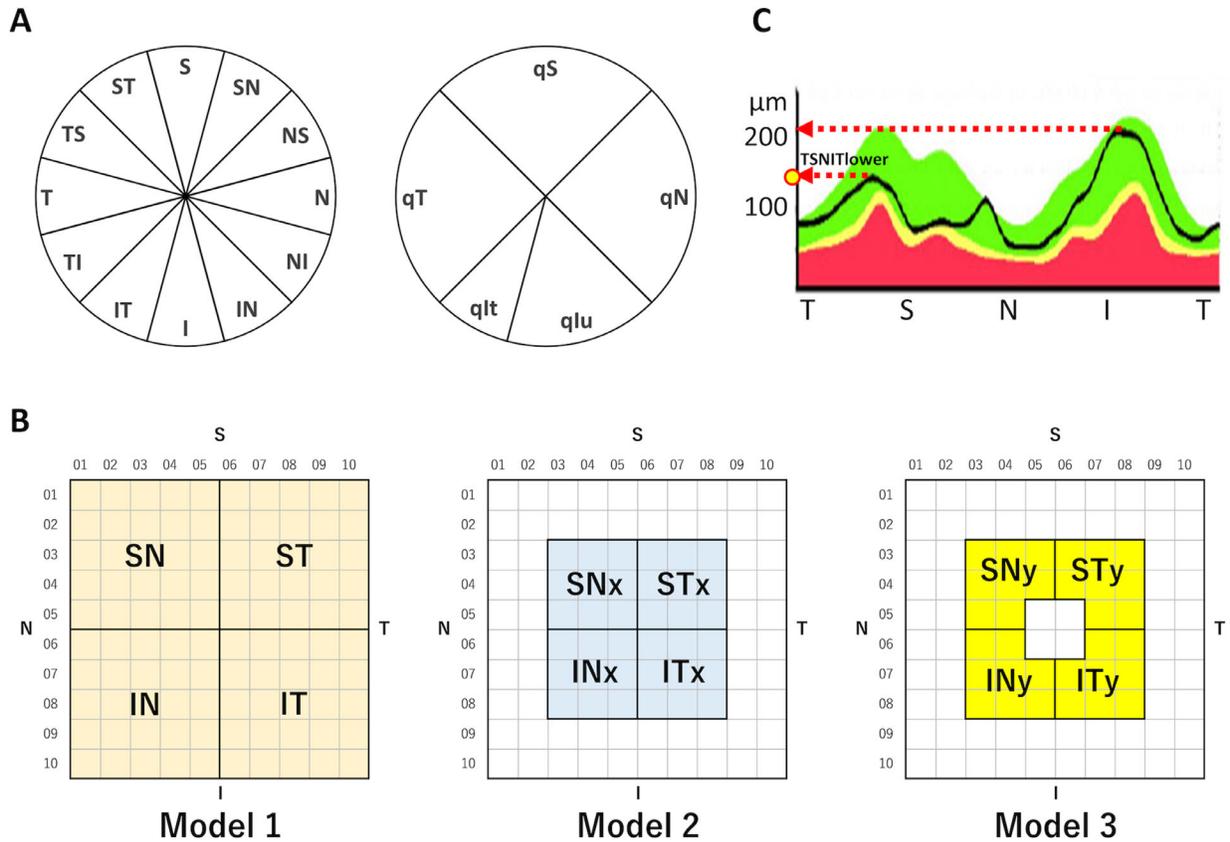
Figure 2. Visualization of potential predictors measured by SD-OCT. (**A**) The thickness of the cpRNFL was obtained in 12 segments and then restructured into 5 regions (named cpRNFLqT, cpRNFLqS, cpRNFLqN, cpRNFLqIn, cpRNFLqIt [μm]). (**B**) The thickness of the mRNFL, macular GCL (mGCL), macular inner plexiform layer (mIPL), macular ganglion cell-inner plexiform layer (mGCIPL; mGCL+mIPL), and macular ganglion cell complex (mGCC; mRNFL+mGCL+mIPL) were obtained in 10 × 10 grids. We restructured these into four regions for model 1 (named ST, SN, IT, and IN [μm]), then created variables that excluded the peripheral regions for model 2 (named STx, SNx, ITx, and INx [μm]), and then another variable that excluded the central region for model 3 (named STy, SNy, ITy, and INy [μm]). (**C**) The thickness of the upper and lower spikes in the temporal-superior-nasal-inferior-temporal (TSNIT) plot was obtained and then the thinner side of those were identified (named TSNITlower [μm]).

examined with an HFA using the 24-2 Swedish Interactive Threshold Algorithm–based standard program. The HFA testing results were evaluated by two ophthalmologists (RT and TaNa). The developed scores were then applied to this subset, and their consistency with the HFA-based assessment was verified.

OCT Measurement Predictors

The macular and peripapillary inner retinal layer thickness was obtained using a three-dimensional wide scan protocol. We restructured the segments and grids and created variables as possible predictors among a total of 312 variables, as described in Figures 2A, B, and C. First, we calculated the average retinal thickness of all 12 segments for circumpapillary RNFL (cpRNFL) and all 100 grids for macular RNFL (mRNFL), macular ganglion cell-inner plexiform layer (mGCIPL), and macular ganglion cell complex (mGCC), respectively. Second, for the 12 segments of cpRNFL, we created variables divided into five regions (named cpRNFLqT, cpRNFLqS, cpRNFLqN, cpRNFLqIn, and cpRNFLqIt) (Fig. 2A). We divided the inferior quadrant according to following studies showing that the inferotemporal RNFL thickness is rather beneficial for glaucoma detection.[32,33] Third, for the 10 × 10 grids of the three macular layers (mRNFL, mGCIPL, and mGCC), we initially restructured these into four regions (named ST, SN, IT, and IN for model 1) (Fig. 2B), then created variables that excluded the peripheral regions (named STx, SNx, ITx, and INx for model 2), and then another variable that excluded the central region (named STy, SNy, ITy, and INy for model 3). The variables for all regions were averaged

over the thickness (µm) of the segment. For example,

$$\text{cpRNFLqS} = (\text{cpRNFL\_S} + \text{cpRNFL\_ST}$$
$$+ \text{cpRNFL\_SN})/3 \ (\mu m)$$
$$\text{mGCC\_SN} = (\text{mGCC } 01\_01 + 01\_02 + \cdots$$
$$+ 05\_04 + 05\_05)/25 \ (\mu m)$$

Fourth, to account for the differences in the thickness of each region,[34,35] variables were created regarding the difference between the superior and inferior portions and the temporal and nasal portions. Because these variables would not take a normal distribution and could take zero if they were exactly the same, natural logarithmic transformations were performed after adding one to the absolute values. For example:

$$\log\_\text{mGCIPL\_STvsIT} = \log_e(|\text{mGCIPL\_ST}$$
$$- \text{mGCIPL\_IT}| + 1)$$

Finally, to account for the case in which the difference in cpRNFL between the superior and inferior is large, but the thickness of the thinner part is not critical, we created a variable that refers to the thickness of the lower side of the upper and lower spikes in the TSNIT plot (named TSNITlower) (see Fig. 2C).

Statistical Analysis

All variables from the OCT measurements were handled as continuous variables. We used multivariable logistic regression models with backward stepwise selection with a P value of 0.1 for backward elimination to select the best predictive model in the development dataset. Three models were examined. For model 1, all the cpRNFL-related values and mRNFL-, mGCIPL-, and mGCC-related variables without exclusion of the peripheral or central regions of the macula grids were included as possible predictors. And then, for macula-related values, we used variables that excluded the peripheral regions for model 2 and variables that excluded the central region for model 3. To add, all models included the overall average thickness (cpRNFL, mRNFL, mGCIPL, and mGCC), the region difference–related variables (e.g., log_mGCIPL_STvsIT), and the TSNIT plot graph-spike-related variable (TSNITlower). The effectiveness of the analyses was estimated by the area under the receiver operating characteristic curves (AUC-ROC).

For validation, we applied the interceptions and betas obtained from the three regression models in the development step and risk scores using the following equations:

$$\log_e\left(\frac{p}{1-p}\right) = Y \Leftrightarrow p = \frac{\exp(Y)}{1 + \exp(Y)} \quad (1)$$

$$\text{Glaucoma screening score} = \text{round}\,(p \times 100), \quad (2)$$

where round(X) denotes the whole number closest to X. In brief, all participants with OCT measurements for each eye took scores from 0 to 100, with higher scores indicating a greater possibility of glaucoma. Based on the score distribution, we divided the population into three groups: 0 to 49, low; 50 to 89, middle; and 90 to 100, high. We then calculated the positive predictive values (PPV) for screening for the necessity of further full ophthalmological examination in each group (PPV for screening). Additionally, sensitivity and specificity were evaluated using diagnoses obtained with subsets additionally augmented with HFA testing.

To clarify the effect of the axial length, we examined the improvement effect of information on the axial length on the risk score. In addition, to confirm accuracy in distinguishing retinal thinning owing to glaucoma from that, owing to high myopia, we estimated axial length from the angle of the double hump of the peripapillary RNFL (Supplementary Fig. S1). First, the raw value of axial length was independently added to model 3. Next, axial length was added to all three models, and the selection of variables in the logistic regression as above was performed. All statistical analyses were performed using SAS 9.4 (SAS, Institute, Cary, NC).

Results

The characteristics of cases and controls in the development data are shown in Table 1. All four layers were significantly thinner in the cases than in the normal controls (overall average of the 10 × 10 grids). Spearman's correlation coefficient of cpRNFL thickness between the left and right eyes was 0.85 ($P < 0.01$) in the control group and 0.24 ($P < 0.01$) in the case group. The distribution by cases and controls for the predictors are illustrated in Supplementary Figure S2.

The diagnostic models developed for glaucoma screening are shown in Table 2. After the selection of variables in the logistic regression models, for the predictors of each region's thickness (cpRNFLqS, cpRNFLqIt, cpRNFLqIn, mGCIPL_IT, mGCC, and TSNITlower), in principle, we obtained negative regression coefficients, meaning a lower possibility of glaucoma if the layers of the retina in the selected area were thicker. In contrast, for the predictors of differences between superior

Table 1. Characteristics of the Participants With Glaucoma Cases and Sex- and Age-matched Controls (Fiscal 2016)

	Cases[a] (284 Eyes of 191 Participants)	Controls[b] (284 Eyes of 277 Participants)	P Value[c]
Gender			1.00
Female	33 (11.6)	33 (11.6)	
Male	251 (88.4)	251 (88.4)	
Age (years)	56.3 ± 9.2	56.2 ± 9.0	.96
SD-OCT measured thickness (μm)			
cpRNFL (10 × 10 grids)	73.3 ± 15.4	99.3 ± 10.1	<.01
cpRNFLqS	88.2 ± 24.3)	123.9 ± 16.5	<.01
cpRNFLqIt	91.5 ± 35.8	142.7 ± 23.6	<.01
cpRNFLqIn	82.4 ± 25.7	119.1 ± 19.3	<.01
TSNITlower	94.8 ± 23.3	140.4 ± 17.8	<.01
mRNFL (10 × 10 grids)	82.0 ± 27.2	123.4 ± 17.8	<.01
mGCIPL (10 × 10 grids)	69.9 ± 11.9	79.5 ± 9.2	<.01
mGCIPL_IT	52.0 ± 9.0	65.3 ± 5.3	<.01
log_mGCIPL_STvsIT	1.9 ± 0.8	0.9 ± 0.6	<.01
log_mGCIPL_ITvsIN	1.7 ± 0.7	1.1 ± 0.6	<.01
log_mGCIPL_STxvsITx	2.1 ± 1.0	1.1 ± 0.6	<.01
log_mGCIPL_SNyvsINy	2.2 ± 1.0	1.1 ± 0.6	<.01
log_mGCIPL_STyvsITy	1.7 ± 0.8	1.2 ± 0.6	<.01
mGCC (10 × 10 grids)	151.9 ± 31.9	202.9 ± 21.2	<.01
mGCC_IN	95.7 ± 15.6	114.8 ± 10.4	<.01
log_mGCC_ITxvsINx	2.3 ± 0.8	1.7 ± 0.6	<.01
log_mGCC_ITyvsINy	2.5 ± 0.8	2.0 ± 0.6	<.01
Axial length (mm)	25.93 ± 1.71	24.43 ± 1.36	<.01

cpRNFL, circumpapillary retinal nerve fiber layer; cpRNFLqIn, inferior-nasal quadrant of cpRNFL; cpRNFLqIt, inferior-temporal quadrant of cpRNFL; cpRNFLqS, superior quadrant of cpRNFL; log_mGCIPL_ITvsIN, log transformed difference between inferior-temporal and inferior-nasal quadrant of mGCIPL (model 1); log_mGCIPL_SNyvsINy, log transformed difference between superior-nasal and inferior-nasal quadrant of mGCIPL (model 3); log_mGCIPL_STxvsITx, log transformed difference between superior-temporal and inferior-temporal quadrant of mGCIPL (model 2); log_mGCIPL_STvsIT, log transformed difference between superior-temporal and inferior-temporal quadrant of mGCIPL (model 1); log_mGCIPL_STyvsITy, log transformed difference between superior-temporal and inferior-temporal quadrant of mGCIPL (model 3); log_mGCC_ITxvsINx, log transformed difference between inferior-temporal and inferior-nasal quadrant of mGCC (model 2); log_mGCC_ITyvsINy, log transformed difference between inferior-temporal and inferior-nasal quadrant of mGCC (model 3); mGCC, macular ganglion cell complex; mGCC_IN, inferior-nasal quadrant of mGCC (model 1); mGCIPL, macular ganglion cell-inner plexiform layer; mGCIPL_IT, inferior-temporal quadrant of mGCIPL; mRNFL, macular retinal nerve fiber layer; SD-OCT, spectral domain optical coherence *tomography;* TSNIT, lower spike in the temporal-superior-nasal-inferior-temporal plot.

Values are number (%) or mean ± standard deviation.

[a]Defined as glaucoma based on complete ophthalmologic examination.

[b]Defined as no findings on either fundus photography or frequency doubling technology perimetry tests by an ophthalmologist.

[c]χ2 test for gender, and *t* test for continuous variables.

and inferior or between temporal and nasal (log_mGCIPL_STvsIT, log_mGCIPL_ITvsIN, log_ mGCIPL_STxvsITx, log_mGCIPL_ITxvsINx, log_mGCC_INyvsITy, log_mGCIPL_SNyvsINy, and log_mGCIPL_STyvsITy), we obtained positive regression coefficients, meaning a higher possibility of glaucoma if the differences between them were larger. Predictor values of the areas under the receiver operating characteristic curve for the three models were all 0.97, showing almost no difference between them. The ROC curves are shown in Supplementary Figure S3.

Table 2. Details of the Three Models Developed for Glaucoma Screening (Fiscal 2016)

Intercept and Retinal Thickness-Related Predictors	B	SE	OR	95% CI	P Value
Model 1					
Intercept	14.4305	2.1692	–	–	<.01
TSNITlower	−0.0404	0.0153	0.96	0.93 −0.99	.01
cpRNFLqS	−0.0303	0.0127	0.97	0.95 −0.99	.02
cpRNFLqIt	−0.0304	0.0097	0.97	0.95 −0.99	<.01
cpRNFLqIn	−0.0271	0.0096	0.97	0.96 −0.99	<.01
mGCIPL_IT	−0.1424	0.0457	0.87	0.79 −0.95	<.01
log_mGCIPL_STvsIT	1.1427	0.2408	3.14	1.96 −5.03	<.01
log_mGCIPL_ITvsIN	0.6971	0.2968	2.01	1.12 −3.59	.02
mGCC_IN	0.0602	0.0270	1.06	1.01 −1.12	.03
AUC-ROC (95%CI)	0.971 (0.960−0.983)				
Model 2					
Intercept	12.6694	1.5639	–	–	<.01
TSNITlower	−0.0329	0.0150	0.97	0.94 −1.00	.03
cpRNFLqS	−0.0344	0.0126	0.97	0.94 −0.99	.01
cpRNFLqIt	−0.0318	0.0078	0.97	0.95 −0.98	<.01
cpRNFLqIn	−0.0398	0.0091	0.96	0.94 −0.98	<.01
log_mGCIPL_STxvsITx	1.1646	0.2164	3.21	2.10 −4.90	<.01
log_mGCC_ITxvsINx	0.7120	0.2425	2.04	1.27 −3.28	<.01
AUC-ROC (95%CI)	0.969 (0.957–0.982)				
Model 3					
Intercept	12.5935	1.7128	–	–	<.01
TSNITlower	−0.0660	0.0108	0.94	0.92 −0.96	<.01
cpRNFLqIn	−0.0296	0.0087	0.97	0.95 −0.99	<.01
mGCC	−0.0289	0.0080	0.97	0.96 −0.99	<.01
log_mGCIPL_SNyvsINy	0.9845	0.2073	2.68	1.78 −4.02	<.01
log_mGCIPL_STyvsITy	0.5671	0.2719	1.76	1.04 −3.00	.04
log_mGCC_ITyvsINy	0.6613	0.2450	1.94	1.20 −3.13	.01
AUC-ROC (95% CI)	0.968 (0.955–0.981)				

AUC-ROC, area under the receiver operating characteristic curve; CI, confidence interval; OR, odds ratio; SE, standard error.

We applied the glaucoma screening scores to the full validation dataset (FY2018; 9720 eyes of 6006 participants). The glaucoma screening scores were calculated using formula (1) and (2) described in the Methods. Specifically, we used the results from the development step to calculate the scores for models 1, 2, and 3, respectively, using Y calculated by the following equations:

Y (model 1)

$$= 14.4305 - 0.0404 \times \text{TSNITlower} - 0.0303$$
$$\times \text{cpRNFLqS} - 0.0304 \times \text{cpRNFLqIt} - 0.0271$$
$$\times \text{cpRNFLqIn} - 0.1424 \times \text{mGCIPL\_IT} + 1.1427$$
$$\times \text{log\_mGCIPL\_STvsIT} + 0.6971$$
$$\times \text{log\_mGCIPL\_ITvsIN} + 0.0602 \times \text{mGCC\_IN}$$

Y (model 2)

$$= 12.6694 - 0.0329 \times \text{TSNITlower} - 0.0344$$
$$\times \text{cpRNFLqS} - 0.0318 \times \text{cpRNFLqIt}$$
$$- 0.0398 \times \text{cpRNFLqIn} + 1.1646$$
$$\times \text{log\_mGCIPL\_STxvsITx} + 0.712$$
$$\times \text{log\_mGCC\_ITxvsINx}$$

Y (model 3)

$$= 12.5935 - 0.066 \times \text{TSNITlower} - 0.0296$$
$$\times \text{cpRNFLqIn} - 0.0289 \times \text{mGCC} + 0.9845$$
$$\times \text{log\_mGCIPL\_SNyvsSTy} + 0.5671$$
$$\times \text{log\_mGCIPL\_STyvsITy} + 0.6613$$
$$\times \text{log\_mGCC\_ITyvsINy}$$

Table 3. Distribution of Glaucoma Screening Scores for the Three Models Developed and PPV of the Need for Further Full Ophthalmologic Examination in the Validation Dataset (Fiscal Y2018)

		Validation Data Set		SD-OCT Report Diagnosed by Four Ophthalmologists		
Model	Risk Score	*N*	%	Total, *N*	Glaucoma, *N*	PPV, %
Model 1						
Low	0 to 49	8380	86.2	157	19	12.1
Middle	50 to 89	750	7.7	183	110	60.1
High	90 to 100	590	6.1	383	309	80.7
Model 2						
Low	0 to 49	8339	85.8	155	18	11.6
Middle	50 to 89	776	8.0	167	86	51.5
High	90 to 100	605	6.2	401	334	83.3
Model 3						
Low	0 to 49	8280	85.2	186	22	11.8
Middle	50 to 89	842	8.7	168	81	48.2
High	90 to 100	598	6.2	369	335	90.8
Total		9720	100	723	438	60.6

PPV, positive prediction value; SD-OCT, spectral domain optical coherence tomography.

Table 4. Results of SD-OCT and HFA Tests Performed in Fiscal 2020 on Patients Suspected to Have Glaucoma Based on Fundus Photography in Fiscal 2016 (*n* = 129) and Comparison With Glaucoma Screening Score

	Model 3 Risk Score		
Diagnosis	High (90 to 100), *N*	Middle (50 to 89), *N*	Low (0 to 49), *N*
All, *N*	67	17	45
Glaucoma	53	8	1
Other eye disease	0	3	4
Undeterminable	10	1	3
Normal	4	5	37

HFA, Humphrey visual field analyzer; SD-OCT, spectral domain optical coherence tomography.

The distribution of the glaucoma screening scores for the three models are shown in Table 3 (see Supplementary Table S1 for detailed scores). Based on models 1, 2, and 3, 6.1%, 6.2%, and 6.2% of the whole population, respectively, corresponding to the high score group ($\geq$90 points). Among the total subjects, we randomly selected two-thirds of the high group, one-fifth of the middle group, and 2% of the low group, with consideration to the time resource of the experts, and assigned four ophthalmologists to judge the OCT reports. Based on models 1, 2, and 3, PPV for screening was 80.7%, 83.3%, and 90.8% in the high group, and negative predictive value was likewise 87.9%, 88.4%, and 88.2% in the low group, respectively, showing model 3 to be the most suitable for glaucoma screening accuracy (Table 3).

The results of the additional validation analysis are shown in Table 4. Based on model 3, which was calculated using the OCT data taken in FY2016, 67

Table 5. Sensitivity and Specificity of the Risk Score for Glaucoma Screening

Model 3 Risk Score	Glaucoma	Normal	Total
90 to 100	53 (85%)	4[*]	57
0 to 89	9[*]	42 (91%)	51
Total	62	46	108

[*]Detailed findings among false-positive and false-negative cases are available in Supplementary Table S2.

subjects (eyes) corresponded with a high score ($\geq$90). Among them, 53 were diagnosed with glaucoma, and 10 could not be determined by OCT and HFA tests, and 4 had no findings (normal). After excluding those with other diseases, sensitivity was 85.4%, and specificity was 91.3% (Table 5). Detailed findings of OCT and HFA in the false-positive and false-negative cases are shown in Supplementary Table S2.

False-positive cases demonstrated suggestive preperimetoric glaucoma (PPG) in two cases and failed to track the optic nerve during OCT scannings in two cases. False-negative cases showed mild RNFL defect (NFLD) with a mild degree of thinning in two cases, narrow NFLD in two cases, and diffuse thinning of the retinal layer owing to myopia in three cases. Two cases were diagnosed with glaucoma; however, those scores based on model 3 were 89 and 88, respectively. Spearman's correlation coefficient between the model 3 risk score and mean deviation and pattern standard deviation of the HFA test was −0.503 (P < 0.001) and 0.618 (P < 0.001). In the cases with a high-risk score and definitive glaucoma, mean deviation and pattern standard deviation (mean ± standard deviation) were −6.80 ± 5.91 and 6.42 ± 4.64, respectively. In nine cases with a low or middle-risk score (<90) and definitive glaucoma (false-negative cases), mean deviation and pattern standard deviation (mean ± standard deviation) were −1.96 ± 1.53 and 3.02 ± 0.76, respectively. There were significant differences between the two groups by the t test.

We also examined the improvement effect of the axial length on the risk score. First, the raw value of axial length was independently added to model 3. The results are shown in Supplementary Table S3a. The odds ratio (95% confidence interval [CI]) of axial length was 1.30 (95% CI, 1.02–1.65). However, accuracy was not improved (Supplementary Tables S3b and S3c). Second, axial length was added to OCT variables, and the selection of variables in the logistic regression was performed in models 1, 2, and 3. The results are shown in Supplementary Table S4a. Odds ratios for axial length were 1.52 (95% CI, 1.21–1.91), 1.56 (95% CI, 1.32–2.10), and 1.52 (95% CI, 1.20–1.93), respectively. Again, however, accuracy was not improved (Supplementary Tables S4b and S4c). In addition, the estimated axial length angle of the double hump of the peripapillary RNFL was not selected as a significant variable in any of the three models (data not shown).

Discussion

In this study, we developed an SD-OCT–based glaucoma risk score for population-based glaucoma mass screening. Development was conducted with reference to the diagnostic logic of glaucoma specialists. This logic included (1) the absolute value of each layer thickness of the whole or separate regions,[36] (2) vertical difference in peripapillary RNFL thickness,[35] (3) vertical and lateral difference of each layer in the macular area,[34] (4) wedge-shaped localized retinal NFLD around the optic disc,[37] (5) difference in the double hump pattern of the peripapillary RNFL,[38] (6) absolute value of the low value of the double hump pattern of the peripapillary RNFL,[38] and (7) axial length estimated from the angle of the double hump pattern of the peripapillary RNFL. Furthermore, in the macula, three patterns were examined using (A) whole data (model 1); (B) exclusion of surrounding areas to avoid the influence of large vessels (model 2)[39]; and (C) exclusion of the central area on the basis that this area is not affected in early stage glaucoma (model 3). On implementation of all three algorithms, the odds ratio of variables regarding the vertical and laterality difference of each layer was the strongest indicator of glaucoma. Model 3 had the highest sensitivity and specificity in the data of SD-OCT alone.

In the validation step, we applied two steps, because of two objectives: one was to develop a score that automatically detects glaucomatous findings in OCT, and the other goal was to screen for glaucoma. Therefore, the first validation was performed by reading OCT by glaucoma specialists, and the second validation was performed by evaluating the glaucoma diagnosis using a Humphrey visual field analyzer. We selected cases in which the risk scores were evenly divided and then added cases that were difficult to diagnose as glaucomatous because the accurate judgment of slight changes in early stage glaucoma is critical for mass screening. This validation process assumed the situation in actual population-based mass screening in which an ophthalmologist makes judgments on the need for detailed examination for glaucoma in hospital. Additionally, the decisions were made by agreement among all four ophthalmologists, blinded to the scores, for the purpose of screening for glaucoma. To validate the accuracy of the risk score using real-world data, cases with low image quality owing to poor fixation, blinking, movement, and small pupils were intentionally included. In some cases, the presence of artifacts, segmentation errors, and optic nerve tracking errors rendered glaucomatous findings difficult to capture; nevertheless, all four ophthalmologists examined the data until they reached a consensus, and then made a final decision. The score was well-related to the judgment needed for a detailed examination for glaucoma. Additionally, we investigated the validity of the scores using ophthalmologic diagnoses based on the HFA test and OCT reports, which confirmed their high performance and accuracy, with a sensitivity of 85.4% and specificity of 91.3%. Four cases showed a high-risk score with no visual field abnormalities and were regarded as false-positive cases. Ophthalmologists judged that some of the false-positives cases were most likely PPG with NFLD detected by OCT. Although

our goal in this study was to detect early and moderate stages of glaucoma, but not to detect PPG, these cases suggest that the risk score could also be used to identify PPG. Nine glaucoma cases showed low scores that were false negatives. The findings from these cases indicated that the risk score tended to be low for cases with a mild thinning type of RNFL, a narrow retinal NFLD, or diffuse retinal layer thinning owing to myopia.

Algorithm-based artificial intelligence is now a mainstream approach,[40] and visualization of deep learning models is an active and ongoing area of research.[21,41] Currently, however, no practical applications have yet appeared because the algorithm is a black box and implementation in the field of mass screening is complex. The algorithm used in this study can be configured to provide a score by the installation of simple calculation software on existing OCT machines, and therefore seems to be suitable for practical use. Furthermore, the cutoff value can be determined according to prevalence in the target population. In this study, we set a score of 90 points or more as requiring detailed examination; the score showed a high PPV for screening; 6% of the 9000 data sets were targeted for a 5% prevalence of glaucoma.[3] Two cases with glaucoma in the additional validation set were false negatives; however, the scores were 88 and 89 points, so the cutoff value could be decreased slightly to increase sensitivity. We believe that this parsimonious model is feasible and reliable, given that the most crucial feature of mass screening is high PPV of screening.

We have developed an OCT screening algorithm that is suitable for screening use. Considering the characteristics of glaucoma, it is essential to observe changes over time. Therefore, it is more appropriate to express our findings in terms of risk score rather than through the setting of a clear cutoff value. The reliability of data based on OCT images owing to deformation of the eyeball by the long axial length, which occurs with severe myopia, should be further considered. High myopia makes differentiation from glaucoma difficult. Independently of glaucoma, the distribution of RNFL thickness will vary considerably depending on the axial length of the eye. We tried modifying the score using an estimation of axial ocular length based on the angle of the TSNIT double hump, but this did not improve accuracy. In fact, even the addition of raw data for axial length did not improve accuracy.

Our study has several strengths. First, we used large-scale data derived from ophthalmologist interpretations of approximately 1291 images as an analysis target. No previous study has evaluated OCT at this scale. Second, all data were population-based and had no hospital bias. Third, this algorithm can explain the reason why a diagnosis of glaucoma is made ophthalmologically rather than as a black box pronouncement. This practice, in turn, aids in convincing hesitant examinees of the need for detailed examination and treatment. Fourth, this study was conducted at our workplace, facilitating follow-up. Of importance considering the progression of glaucoma, the final validation was made based on results four years after the initial diagnosis.

Small limitations also warrant mention. The number of validation subjects with a final diagnosis was limited. However, our primary goal was not a final diagnosis of glaucoma, but rather to develop an algorithm to aid ophthalmologists in determining whether an examinee requires a detailed hospital evaluation. For this purpose, our validation was sufficient.

In conclusion, the SD-OCT-based glaucoma risk score developed in this study can be used in mass glaucoma screening and will play an important role in this screening.

References

1. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol*. 2006;90(3):262–267.
2. Tham Y-C, Li X, Wong TY, Quigley HA, Aung T, Cheng C-Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040. *Ophthalmology*. 2014;121(11):2081–2090.

translational vision science & technology

3. Iwase A, Suzuki Y, Araie M, et al. The prevalence of primary open-angle glaucoma in Japanese: the Tajimi Study. *Ophthalmology.* 2004;111(9):1641–1648.

4. Shen SY, Wong TY, Foster PJ, et al. The prevalence and types of glaucoma in Malay people: the Singapore Malay eye study. *Invest Ophthalmol Vis Sci.* 2008;49(9):3846–3851.

5. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review. *JAMA.* 2014;311(18):1901–1911.

6. Tatemichi M, Nakano T, Tanaka K, et al. Performance of glaucoma mass screening with only a visual field test using frequency-doubling technology perimetry. *Am J Ophthalmol.* 2002;134(4):529–537.

7. Tatemichi M, Nakano T, Tanaka K, et al. Laterality of the performance of glaucoma mass screening using frequency-doubling technology. *J Glaucoma.* 2003;12(3):221–225.

8. Yamada M, Hiratsuka Y, Nakano T, et al. Detection of glaucoma and other vision-threatening ocular diseases in the population recruited at specific health checkups in Japan. *Clin Epidemiol.* 2020;12:1381–1388.

9. Terauchi R, Wada T, Ogawa S, et al. FDT perimetry for glaucoma detection in comprehensive health checkup service. *J Ophthalmol.* 2020;2020:1–6.

10. Tang J, Liang Y, O'Neill C, Kee F, Jiang J, Congdon N. Cost-effectiveness and cost-utility of population-based glaucoma screening in China: a decision-analytic Markov model. *Lancet Glob. Health.* 2019;7(7):e968–e978.

11. John D, Parikh R. Cost-effectiveness of community screening for glaucoma in rural India: a decision analytical model. *Public Health.* 2018;155:142–151.

12. Varma R, Lee PP, Goldberg I, Kotak S. An assessment of the health and economic burdens of glaucoma. *Am J Ophthalmol.* 2011;152(4):515–522.

13. Tan NYQ, Friedman DS, Stalmans I, Ahmed IIK, Sng CCA. Glaucoma screening: where are we and where do we need to go? *Curr Opin Ophthalmol.* 2020;31(2):91–100.

14. Stein JD, Kim DS, Niziol LM, et al. Differences in rates of glaucoma among Asian Americans and other racial groups, and among various Asian ethnic groups. *Ophthalmology.* 2011;118(6):1031–1037.

15. Stein JD, Khawaja AP, Weizer JS. Glaucoma in adults—screening, diagnosis, and management. *JAMA.* 2021;325(2):164–174.

16. Bourne RR. Worldwide glaucoma through the looking glass. *Br J Ophthalmol.* 2006;90(3):253–254.

17. Fujimoto JG, Pitris C, Boppart SA, Brezinski ME. Optical coherence tomography: An emerging technology for biomedical imaging and optical biopsy. *Neoplasia.* 2000;2(1-2):9–25.

18. Hee MR, Izatt JA, Swanson EA, et al. Optical coherence tomography of the human retina. *Arch Ophthalmol.* 1995;113(3):325–332.

19. Huang D, Swanson E, Lin C, et al. Optical coherence tomography. *Science.* 1991;254(5035):1178–1181.

20. Banister K, Boachie C, Bourne R, et al. Can automated imaging for optic disc and retinal nerve fiber layer analysis aid glaucoma detection? *Ophthalmology.* 2016;123(5):930–938.

21. Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Transl Vis Sci Technol.* 2020;9(2):42.

22. Nakano T, Hayashi T, Nakagawa T, et al. Applicability of automatic spectral domain optical coherence tomography for glaucoma mass screening. *Clin Ophthalmol.* 2016;11:97–103.

23. Mayama C, Saito H, Hirasawa H, et al. Diagnosis of early-stage glaucoma by grid-wise macular inner retinal layer thickness measurement and effect of compensation of disc-fovea inclination. *Invest Ophthalmol Vis Sci.* 2015;56(9):5681–5690.

24. Yoshida T, Iwase A, Hirasawa H, et al. Discriminating between glaucoma and normal eyes using optical coherence tomography and the 'Random Forests' classifier. *PLoS One.* 2014;9(8):e106117.

25. Zheng C, Johnson TV, Garg A, Boland MV. Artificial intelligence in glaucoma. *Curr Opin Ophthalmol.* 2019;30(2):97–103.

26. Medeiros FA. Deep learning in glaucoma: progress, but still lots to do. *Lancet Digital Health.* 2019;1(4):e151–e152.

27. Ran AR, Tham CC, Chan PP, et al. Deep learning in glaucoma with optical coherence tomography: a review. *Eye.* 2021;35(1):188–201.

28. Mirzania D, Thompson AC, Muir KW. Applications of deep learning in detection of glaucoma: a systematic review. *Eur J Ophthalmol.* 2021;31(4):1618–1642.[published online December 4, 2020].

29. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA.* 2013;310(20):2191–2194.

30. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable predic-

tion model for Individual Prognosis Or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1–W1.

31. Honda T, Nakagawa T, Watanabe Y, et al. Association between information and communication technology use and ocular axial length elongation among middle-aged male workers. *Sci Rep.* 2019;9(1):17489.

32. Budenz DL, Michael A, Chang RT, McSoley J, Katz J. Sensitivity and specificity of the StratusOCT for perimetric glaucoma. *Ophthalmology.* 2005;112(1):3–9.

33. Leung CK-S, Chan W-M, Chong KK-L, et al. Comparative study of retinal nerve fiber layer measurement by StratusOCT and GDx VCC, I: correlation analysis in glaucoma. *Invest Ophthalmol Vis Sci.* 2005;46(9):3214–3220.

34. Um TW, Sung KR, Wollstein G, Yun S-C, Na JH, Schuman JS. Asymmetry in hemifield macular thickness as an early indicator of glaucomatous change. *Invest Ophthalmol Vis Sci.* 2012;53(3):1139–1144.

35. Budenz D, Michael A, Chang R, McSoley J, Katz J. Sensitivity and specificity of the StratusOCT for perimetric glaucoma. *Ophthalmology.* 2005;112(1):3–9.

36. Medeiros FA, Zangwill LM, Bowd C, Vessani RM, Susanna R, Weinreb RN. Evaluation of retinal nerve fiber layer, optic nerve head, and macular thickness measurements for glaucoma detection using optical coherence tomography. *Am J Ophthalmol.* 2005;139(1):44–55.

37. Jonas JB, Schiro D. Localised wedge shaped defects of the retinal nerve fibre layer in glaucoma. *Br J Ophthalmol.* 1994;78(4):285–290.

38. Park J-W, Jung H-H, Heo H, Park S-W. Validity of the temporal-to-nasal macular ganglion cell-inner plexiform layer thickness ratio as a diagnostic parameter in early glaucoma. *Acta Ophthalmol.* 2015;93(5):e356–e365.

39. Takemoto D, Higashide T, Ohkubo S, Udagawa S, Sugiyama K. Ability of macular inner retinal layer thickness asymmetry evaluated by optical coherence tomography to detect preperimetric glaucoma. *Transl Vis Sci Technol.* 2020;9(5):8.

40. Seo SB, Cho H-K. Deep learning classification of early normal-tension glaucoma and glaucoma suspects using Bruch's membrane opening-minimum rim width and RNFL. *Sci Rep.* 2020;10(1):19042.

41. Thakur A, Goldbaum M, Yousefi S. Predicting glaucoma before onset using deep learning. *Ophthalmol Glaucoma.* 2020;3(4):262–268.