AMERICAN ACADEMY OF OPHTHALMOLOGY®

Deep Imbalanced Regression Model for Predicting Refractive Error from Retinal Photos

Samantha Min Er Yew, BSc,[1,2,*] *Xiaofeng Lei, MSc,*[3,*] *Yibing Chen, BEng,*[4] *Jocelyn Hui Lin Goh, BEng,*[4] *Krithi Pushpanathan, MSc,*[1,2] *Can Can Xue, MD, PhD,*[4] *Ya Xing Wang, MD, PhD,*[5] *Jost B. Jonas, MD, PhD,*[6,7] *Charumathi Sabanayagam, MD, PhD,*[4,8] *Victor Teck Chang Koh, MBBS, MMed,*[1,2] *Xinxing Xu, PhD,*[3] *Yong Liu, PhD,*[3] *Ching-Yu Cheng, MD, PhD,*[1,2,4,8,†] *Yih-Chung Tham, PhD*[1,2,4,8,†,]

Purpose: Recent studies utilized ocular images and deep learning (DL) to predict refractive error and yielded notable results. However, most studies did not address biases from imbalanced datasets or conduct external validations. To address these gaps, this study aimed to integrate the deep imbalanced regression (DIR) technique into ResNet and Vision Transformer models to predict refractive error from retinal photographs.

Design: Retrospective study.

Subjects: We developed the DL models using up to 103 865 images from the Singapore Epidemiology of Eye Diseases Study and the United Kingdom Biobank, with internal testing on up to 8067 images. External testing was conducted on 7043 images from the Singapore Prospective Study and 5539 images from the Beijing Eye Study. Retinal images and corresponding refractive error data were extracted.

Methods: This retrospective study developed regression-based models, including ResNet34 with DIR, and SwinV2 (Swin Transformer) with DIR, incorporating Label Distribution Smoothing and Feature Distribution Smoothing. These models were compared against their baseline versions, ResNet34 and SwinV2, in predicting spherical and spherical equivalent (SE) power.

Main Outcome Measures: Mean absolute error (MAE) and coefficient of determination were used to evaluate the models' performances. The Wilcoxon signed-rank test was performed to assess statistical significance between DIR-integrated models and their baseline versions.

Results: For prediction of the spherical power, ResNet34 with DIR (MAE: 0.84D) and SwinV2 with DIR (MAE: 0.77D) significantly outperformed their baseline—ResNet34 (MAE: 0.88D; $P < 0.001$) and SwinV2 (MAE: 0.87D; $P < 0.001$) in internal test. For prediction of the SE power, ResNet34 with DIR (MAE: 0.78D) and SwinV2 with DIR (MAE: 0.75D) consistently significantly outperformed its baseline—ResNet34 (MAE: 0.81D; $P < 0.001$) and SwinV2 (MAE: 0.78D; $P < 0.05$) in internal test. Similar trends were observed in external test sets for both spherical and SE power prediction.

Conclusions: Deep imbalanced regressed–integrated DL models showed potential in addressing data imbalances and improving the prediction of refractive error. These findings highlight the potential utility of combining DL models with retinal imaging for opportunistic screening of refractive errors, particularly in settings where retinal cameras are already in use.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2025;5:100659 © 2024 by the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Uncorrected refractive error is the leading cause of moderate and severe visual impairment (VI) in adults >50 years old, accounting for 86.1% of cases.[1] Uncorrected refractive error can have a profound impact on quality of life and work productivity.[2] Despite the cost-effectiveness of spectacle correction, only 35.7% of individuals with distance-related uncorrected refractive error receive treatment globally.[3] This situation highlights a significant public health issue and underscores the critical need for effective global interventions.

Advancements in deep learning (DL) demonstrated promising performance in ophthalmology[4–11] and position DL as a potential tool to enhance screening effectiveness and thereby reduce VI. To that end, coupling DL algorithms with retinal imaging offers a promising avenue for opportunistic screening of refractive error, especially as retinal

https://doi.org/10.1016/j.xops.2024.100659
ISSN 2666-9145/24

cameras become increasingly available in vision screening programs such as diabetic retinopathy (DR) screening.

Data imbalance is a common and inherent issue in real-world application, where certain target values are represented by significantly fewer observations.[12] Past studies have used DL models for predicting refractive error using ocular images; however, most of the studies did not address biases from imbalanced datasets or conduct external validation.[13–21] Deep learning models trained on imbalanced data tend to produce biased results and may not generalize well to external test sets.[22,23] Therefore, addressing data imbalances is crucial for improving the model's generalizability and clinical applicability.[24]

To tackle this, we integrated the deep imbalanced regression (DIR) technique[12] into 2 DL models, namely—ResNet34[25] and SwinV2[26] (Swin Transformer version 2). In this study, we evaluated and compared the performances of these DIR-integrated models with conventional DL models without DIR in predicting spherical and spherical equivalent (SE) power. We hypothesize that incorporating DIR will improve the accuracy of refractive error predictions.

Methods

Study Design

In this retrospective study, we trained and tested our models (ResNet34, ResNet34 with DIR, SwinV2, and SwinV2 with DIR) using retinal images from the Singapore Epidemiology of Eye Diseases (SEED) Study[27] and the United Kingdom Biobank (UKBB)[28] and externally tested the models on the Singapore Prospective Study Program (SP2)[29] and the Beijing Eye Study (BES)[30] datasets (Fig 1). We extracted the relevant clinical data from the 4 study cohorts. The relevant clinical data obtained were the patients' identification codes, refractive error status (SE and spherical power), and macular centered retinal images. Images of poor quality and those from patients who had undergone cataract surgery were excluded from the data extraction process. Retinal photographs of poor quality or with artefacts due to eye movement, blinking, or extremely small pupil (thus restricted view of the fundus) were excluded. In SEED, BES, and SP2 datasets, image quality was assessed through manual inspection by the original studies' representatives or readers. In UKBB dataset, we employed the RetiSort[31] algorithm to filter and select good quality retinal images.

Written informed consent was obtained from all participants. Each study complied with the principles of the Declaration of Helsinki and received approval from the respective local ethical committees. Additionally, data usage permission was granted by the principal investigator of each study.

Study Dataset

For the prediction of the spherical power, we employed a balanced splitting approach. Using the combined SEED and UKBB dataset of 111 932 retinal images (SEED: 17 105 and UKBB: 94 827), we allocated 95 798 images (SEED: 13 325 and UKBB: 82 473) for training, 8067 images (SEED: 1890 and UKBB: 6177) for validation, and another 8067 images (SEED: 1890 and UKBB: 6177) for internal testing, adhering to an approximate ratio of 12:1:1 (Table S1, available at www.ophthalmologyscience.org).

For the prediction of SE, we employed a balanced splitting approach. Using the combined SEED and UKBB dataset of 111 684 retinal images (SEED: 16 857 and UKBB: 94 827), we allocated 95 670 images (SEED: 13 079 and UKBB: 82 591) for training, 8007 images (SEED: 1889 and UKBB: 6118) for validation, and another 8007 images (SEED: 1889 and UKBB: 6118) for internal testing, adhering to an approximate ratio of 12:1:1 (Table S1, available at www.ophthalmologyscience.org).

For the detection of significant refractive error, we employed a random data splitting approach. Using the combined SEED and UKBB dataset of 111 794 retinal images (SEED: 16 967 and UKBB: 94 827), we allocated 88 037 images (SEED: 13 363 and UKBB: 74 674) for training, 9782 images (SEED: 1515 and UKBB: 8267) for validation, and another 13 975 images (SEED: 2089 and UKBB: 11 886) for internal testing, adhering to an approximate ratio of 8:1:1 (Table S1, available at www.ophthalmologyscience.org).

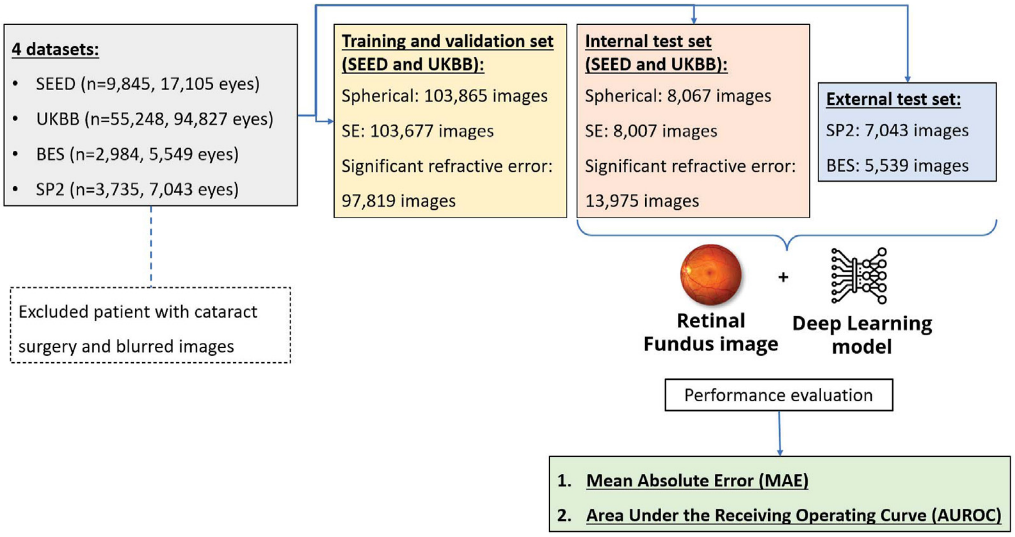
4 datasets:
SEED (n=9,845, 17,105 eyes)
UKBB (n=55,248, 94,827 eyes)
BES (n=2,984, 5,549 eyes)
SP2 (n=3,735, 7,043 eyes)
Excluded patient with cataract surgery and blurred images
Training and validation set (SEED and UKBB):
Spherical: 103,865 images
SE: 103,677 images
Significant refractive error: 97,819 images
Internal test set (SEED and UKBB):
Spherical: 8,067 images
SE: 8,007 images
Significant refractive error: 13,975 images
External test set:
SP2: 7,043 images
BES: 5,539 images
Retinal Fundus image
+
Deep Learning model
Performance evaluation
1. Mean Absolute Error (MAE)
2. Area Under the Receiving Operating Curve (AUROC)

Figure 1. Overview of study design. BES = Beijing Eye Study; SE = spherical equivalent; SEED = Singapore Epidemiology of Eye Diseases; SP2 = Singapore Prospective Study; UKBB = United Kingdom Biobank.

For each of the external testing set in the prediction of spherical power, SE, and significant refractive error, 7043 retinal images from SP2 and 5539 retinal images from BES were used (Table S1, available at www.ophthalmologyscience.org).

DL Model Development

We developed a regression model based on a supervised DL approach to predict refractive error using retinal photographs. For the underlying model architecture, we utilized the ResNet34 and SwinV2 model coupled with DIR technique[12] to minimize the effects of imbalanced data on the model's predictive performance. We then evaluated the performance of the ResNet34 with DIR and SwinV2 with DIR model against the original ResNet34 and SwinV2 model in predicting spherical power and SE.

The integration of DIR was anticipated to enhance the model's performance. The labels were retrieved from 4 datasets with corresponding retinal images. Two simple yet effective methods, label distribution smoothing and feature distribution smoothing (Fig 2; Figs S3 and S4, available at www.ophthalmologyscience.org), were employed to address DIR, exploiting the similarity between nearby targets in both label and feature space. This approach considers the impact of nearby target values and adjusts the distributions of both the labels and learned features. The primary inputs to each of the DL models were the macula-centered retinal images and the relevant clinical labels (i.e., Spherical and SE power). Label distribution smoothing is similar to the reweighting method in classification model, but it considers the difference in classification and regression models (i.e., the empirical label distribution corresponding to the real label density distribution is different in the 2-type model). Label distribution smoothing convolves a symmetric kernel with the empirical label density to estimate the effective label density distribution that accounts for the continuity of labels and captures the real imbalance that affects regression problems. Feature distribution smoothing assumes the feature of the neighbor continuous target values should be similar; it introduces a feature calibration layer that uses kernel smoothing to smooth the distributions of feature mean and covariance over the target space for addressing data imbalance.[30] We applied data augmentation techniques, including random horizontal flipping, scaling, rotation, and ColorJitter. Additionally, the square-root inverse reweighting method was employed with DIR during model training.[12,32]

We also developed binary-based models using ResNet34 and SwinV2 backbone with a supervised DL approach aimed at determining the presence of a significant refractive error (e.g., 'Yes' or 'No' of significant refractive error) (Fig S5, available at www.ophthalmologyscience.org). The positive to negative class ratio is balanced. To the best of our knowledge, to date, there is no standardized definition for significant refractive error. Nevertheless, we have adopted an approach of corresponding the degree of refractive error with significant VI (defined as 20/40, based on United States definition).[33] We defined significant refractive error as a spherical power of ≤ -1.00D or $\geq +1.50$D, and/or a cylindrical power of ≤ -1.50D. The spherical power threshold of −1.00DS was selected because it corresponds to a visual acuity of 20/40. Although +1.00DS theoretically corresponds to 20/40 vision, clinical experience shows that hyperopic patients with +1.00DS often achieve better than 20/40 vision.[34,35] Therefore, we opted for a +1.50DS threshold. The cylinder power threshold of −1.50DC was selected because it corresponds to a visual acuity of 20/40, and patients with this refractive error may experience impaired vision, particularly in low-contrast conditions, increasing the risk of falls.[36,37]

Statistical Analysis

For each retinal image, the regression model generated a continuous prediction (i.e., spherical and SE). Mean absolute error (MAE) and coefficient of determination (R^2) were used to evaluate its performance. Data normality was assessed using the Shapiro−Wilk test. The normality test (Shapiro−Wilk test) for the difference between actual and predicted regression values showed that the distribution of these measurements did not follow a normal distribution. Hence, we performed Wilcoxon signed-rank test to assess the differences between the actual and predicted values across all models.

Agreement between the actual and predicted values was evaluated using Bland−Altman plots.[38,39] The 95% limits of agreement (LOA) were defined as the mean difference ±1.96 standard deviations. The difference [artificial intelligence-predicted power minus ground truth] were plotted against the average of the 2 measurements. When a trend was identified, Pearson's correlation coefficient (R) of the least squares regression line was plotted.

For each retinal image, the classification model generated probability output values (from 0 to 1) which corresponded to the probability of the image having a significant refractive error or a nonsignificant refractive error. Based on the probability scores, we selected an optimal threshold based on the Youden Index to determine the predicted binary class for each retinal image. "0" represented nonsignificant refractive error and "1" represented a significant refractive error. To assess the model's performance for

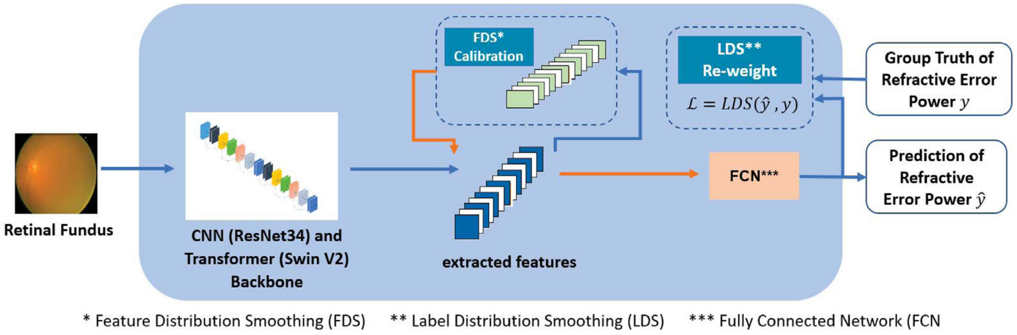

Figure 2. Deep imbalanced regression model using convolutional neural network and transformer backbone for predicting refractive error from retinal fundus image. Label distribution smoothing and feature distribution smoothing, were employed to address deep imbalanced regression, exploiting the similarity between nearby targets in both label and feature space. This approach considers the impact of nearby target values and adjusts the distributions of both the labels and learned features.

binary classification, we used the area under the receiver operating characteristics curve (AUROC).

Statistical analyses were performed using Jupyter Lab, Python 3.11.5, packaged by Anaconda, Inc. (main, Sep 11 2023, 13:26:23, MSC v.1916 64 bit [AMD64]).

Results

The characteristics of the 4 datasets (SEED, UKBB, SP2, and BES) were summarized in Table 2. The mean age of subjects in SEED was 58.9 ± 10.4 years, 56.5 ± 8.1 years in UKBB, 49.9 ± 12.6 years in SP2, and 56.2 ± 10.6 years in BES. The SEED dataset comprised 9845 participants (17 105 eyes), and the UKBB dataset comprised 55 248 patients (94 827 eyes). The SP2 dataset comprised 3735 patients (7043 eyes), and the BES dataset comprised 2984 patients (5539 eyes). The proportion of women ranged between 50.6% and 56.2% across the datasets. The SEED dataset included individuals of Malay (n = 3326, 33.8%), Indian (n = 3226, 32.8%), and Chinese (n = 3293, 33.4%) ethnicities. The UKBB dataset included individuals of White (n = 89 033, 94.4%), African ancestry (n = 1725, 1.8%), South Asian (1686, 1.8%), Chinese (n = 263, 0.3%), and other (n = 1593, 1.7%) ethnicities. The SP2 dataset included individuals of Malay (n = 841, 22.5%), Indian (n = 179, 19.3%), Chinese (n = 2172, 58.1%), and other (n = 4, 0.1%) ethnicities. The BES dataset only included Chinese. Retinal images captured in the SEED, SP2, and BES were taken under medical mydriasis, whereas those from the UKBB were obtained with the pupils undilated.

SE Prediction in ResNet34 and ResNet34 with DIR Models

For spherical power prediction in the internal test, ResNet34 with DIR achieved an MAE of 0.84D and R^2 of 0.93 compared with RestNet34's MAE of 0.88D and R^2 of 0.92 ($P < 0.001$). In the SP2 dataset, ResNet34 with DIR achieved an MAE of 0.84D and R^2 of 0.79, compared with ResNet34's MAE of 0.98D and R^2 of 0.74 ($P < 0.001$). In the BES dataset, ResNet34 with DIR achieved an MAE of 0.56 and R^2 of 0.78 compared with ResNet34's MAE of 0.57D and R^2 of 0.77 ($P < 0.001$) (Table 3).

SE Prediction in SwinV2 and SwinV2 with DIR Models

For spherical power prediction in the internal test, SwinV2 with DIR achieved an MAE of 0.77D and R^2 of 0.94, compared with SwinV2's MAE of 0.87D and R^2 of 0.92 ($P < 0.001$). In the SP2 dataset, SwinV2 with DIR achieved an MAE of 0.80 and R^2 of 0.80, compared with SwinV2's MAE of 0.90D and R^2 of 0.76 ($P < 0.001$). In the BES dataset, SwinV2 with DIR achieved an MAE of 0.61 and R^2 of 0.75, compared with SwinV2's MAE of 0.63D and R^2 of 0.73 ($P < 0.001$) (Table 3).

Table 2. Characteristics of Datasets Used

Dataset	SEED (n = 9845)	UKBB (n = 55 248)	SP2 (n = 3735)	BES (n = 2984)
Mean age, yrs (±SD)	58.9 ± 10.4	56.5 ± 8.1	49.9 ± 12.6	56.2 ± 10.6
Ethnicity	Malay n = 3326 (33.8%) Indian n = 3226 (32.8%) Chinese n = 3293 (33.4%)	White n = 89 033 (94.4%) African ancestry n = 1725 (1.8%) South Asian n = 1686 (1.8%) Chinese n = 263 (0.3%) Others n = 1593 (1.7%)	Malay n = 841 (22.5%) Indian n = 179 (19.3%) Chinese n = 2172 (58.1%) Others n = 4 (0.1%)	Chinese n = 2984 (100%)
Number of eyes	17 105	94 827	7043	5539
Gender				
Male	n = 4850 (49.3%)	n = 24 889 (45%)	n = 1792 (48%)	n = 1309 (43.9%)
Female	n = 4995 (50.7%)	n = 30 359 (55%)	n = 1943 (52%)	n = 1675 (56.1%)
Country	Singapore	United Kingdom	Singapore	China
Type of cohort	Population based	Population based	Population based	Population based
Camera used	Canon CR-1 Mark-II NM, Japan	Topcon 3D OCT-1000 Mark II system, Japan	Canon CR-DGi NM, Japan	Canon CR6-45NM, Japan
Field of view	45°	45°	45°	45°
Dilated or undilated fundus	Dilated	Undilated	Dilated	Dilated
Biometry machine	Canon RK-5 Auto Ref-Keratometer	Tomey RC-5000 Auto Ref-Keratometer	Canon RK-5 Auto Ref-Keratometer	Nidek AR-610 Autorefractor

BES = Beijing Eye Study; SD = standard deviation; SEED = Singapore Epidemiology of Eye Diseases; SP2 = Singapore Prospective Study; UKBB = United Kingdom Biobank.

Table 3. Performance Comparison of ResNet34 and ResNet34 with DIR and SwinV2 and SwinV2 with DIR in Spherical Power Prediction

	ResNet34		ResNet34 with DIR			SwinV2		SwinV2 with DIR		
Output: Spherical Power (D)	*MAE*	R^2	*MAE*	R^2	***P* Value***	*MAE*	R^2	*MAE*	R^2	***P* Value**[†]
Internal dataset										
SEED and UKBB (n = 8067 images)	0.88	0.92	0.84	0.93	**<0.001**	0.87	0.92	0.77	0.94	**<0.001**
External datasets										
SP2 (n = 7043 images)	0.98	0.74	0.84	0.79	**<0.001**	0.90	0.76	0.80	0.80	**<0.001**
BES (n = 5539 images)	0.57	0.77	0.56	0.78	**<0.001**	0.63	0.73	0.61	0.75	**<0.001**

BES = Beijing Eye Study; DIR = Deep Imbalanced Regression; MAE = mean absolute error; R^2 = coefficient of determination; SEED = Singapore Epidemiology of Eye Diseases; SP2 = Singapore Prospective Study; UKBB = United Kingdom Biobank.
Bolded *P* value indicates significant difference.
**P* value of <0.05 indicates statistically significantly difference in MAE between ResNet34 with DIR and ResNet34.
[†]*P* value of <0.05 indicates statistically significantly difference in MAE between SwinV2 with DIR and SwinV2.

SE Prediction in ResNet34 and ResNet34 + DIR

For SE prediction in the internal test, ResNet34 with DIR achieved an MAE of 0.78D and R^2 of 0.94 compared with ResNet34's MAE of 0.81D and R^2 of 0.93 ($P < 0.001$). In the SP2 dataset, ResNet34 with DIR achieved an MAE of 0.84 and R^2 of 0.79 compared with ResNet34's MAE of 0.85D and R^2 of 0.78 (P: 0.97). In the BES dataset, ResNet34 with DIR achieved an MAE of 0.51D and R^2 of 0.82 compared with ResNet34's MAE of 0.54D and R^2 of 0.80 ($P < 0.001$) (Table 4).

SE Prediction in SwinV2 and SwinV2 with DIR Models

For SE prediction in the internal test, SwinV2 with DIR achieved an MAE of 0.75D and R^2 of 0.94, compared with SwinV2's MAE of 0.78D and R^2 of 0.93 ($P = 0.001$). In the SP2 dataset, SwinV2 with DIR had an MAE of 0.75D and R^2 of 0.82, compared with SwinV2's MAE of 0.79D and R^2 of 0.81 ($P < 0.001$). In the BES dataset, SwinV2 with DIR showed an MAE of 0.51D and R^2 of 0.82, compared with SwinV2's MAE of 0.64D and R^2 of 0.75 ($P < 0.001$) (Table 4).

Agreement between Predicted Values by DL Model Predicted Values and Ground Truth

Figures 6 and 7 show the Bland−Altman plots illustrating the agreement between actual and predicted power in both spherical and SE prediction for internal validation. Figure 6A shows the Bland−Altman plot of ResNet34 model for spherical power prediction (mean difference: −0.14D; LOA: −2.53, 2.24). Figure 6B shows the Bland−Altman plot of ResNet34 with DIR for spherical prediction (mean difference: −0.06D; LOA: −2.34, 2.21). Figure 6C shows the Bland−Altman plot of SwinV2 model for spherical power prediction (mean difference: 0.11D; LOA: −2.25, 2.47). Figure 6D shows the Bland−Altman plot of SwinV2 with DIR model for spherical power prediction (mean difference: 0.04D; LOA: −2.17, 2.09). Figure 7A shows the Bland−Altman plot of ResNet34 for SE prediction (mean difference: 0.06D; LOA: −2.31, 2.44). Figure 7B shows the Bland−Altman plot of ResNet34 with DIR model for SE prediction (mean difference: 0.06D; LOA: −2.17, 2.30). Figure 7C shows the Bland−Altman plot of SwinV2 model for SE prediction (mean difference: 0.17D; LOA: −2.10, 2.44). Figure 7D shows the Bland−Altman plot of SwinV2

Table 4. Performance Comparison of ResNet34 and ResNet34 with DIR and SwinV2 and SwinV2 with DIR in Spherical Equivalent Prediction

	ResNet34		ResNet34 with DIR			SwinV2		SwinV2 with DIR		
Output: Spherical Equivalent (D)	*MAE*	R^2	*MAE*	R^2	***P* Value***	*MAE*	R^2	*MAE*	R^2	***P* Value**[†]
Internal dataset										
SEED and UKBB (n = 8007 images)	0.81	0.93	0.78	0.94	**<0.001**	0.78	0.93	0.75	0.94	**0.001**
External datasets										
SP2 (n = 7043 images)	0.85	0.78	0.84	0.79	0.97	0.79	0.81	0.75	0.82	**<0.001**
BES (n = 5539 images)	0.54	0.80	0.51	0.82	**<0.001**	0.64	0.75	0.51	0.82	**<0.001**

BES = Beijing Eye Study; DIR = Deep Imbalanced Regression; MAE = mean absolute error; R^2 = coefficient of determination; SEED = Singapore Epidemiology of Eye Diseases; SP2 = Singapore Prospective Study; UKBB = United Kingdom Biobank.
Bolded *P* value indicates significant difference.
**P* value of <0.05 indicates statistically significantly difference in MAE between ResNet34 with DIR and ResNet34.
[†]*P* value of <0.05 indicates statistically significantly difference in MAE between SwinV2 with DIR and SwinV2.

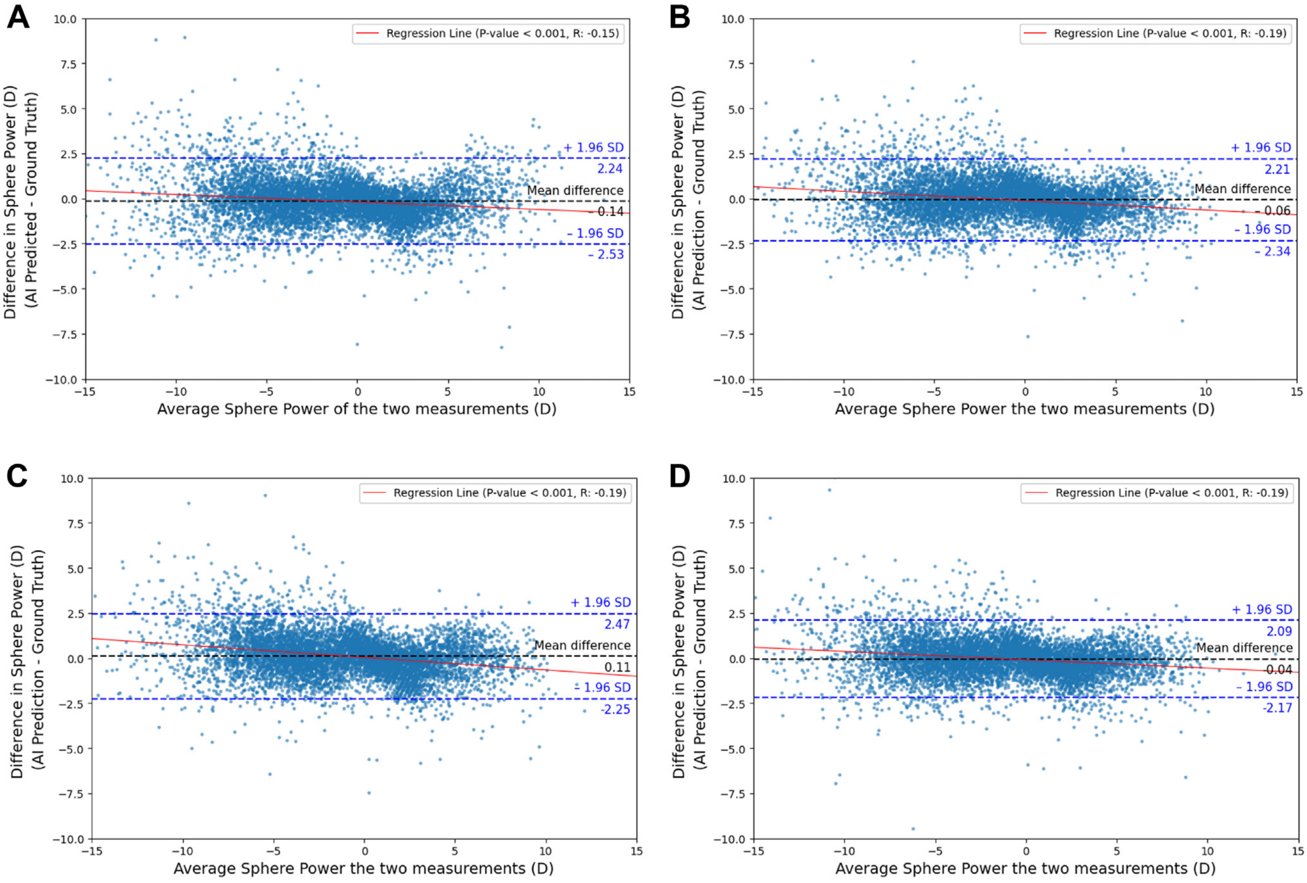

Figure 6. Bland–Altman plots of spherical power measurements in internal test: **A,** model: ResNet34 (MAE: 0.88D, R^2: 0.92); **B,** model: ResNet34 with DIR (MAE: 0.84D, R^2: 0.93); **C,** model: SwinV2 (MAE: 0.87D, R^2: 0.92); **D,** model: SwinV2 with DIR (MAE: 0.77D, R^2: 0.94). Red line represents the least squares regression line. DIR = deep imbalanced regression; MAE = mean absolute error; SD = standard deviation.

with DIR model for SE prediction (mean difference: 0.14D; LOA: −2.04, 2.32). The Bland–Altman plots for the remaining datasets can be found in Figures S8–S11 (available at www.ophthalmologyscience.org).

Significant Refractive Error Prediction with ResNet34 and SwinV2 Model

Using the SwinV2 model for the detection of significant refractive error, the AUROC was 0.90 for the internal test. For external test, the AUROC was 0.87 in the SP2 dataset and 0.79 in the BES dataset. In comparison, the ResNet34 model achieved an AUROC of 0.86 for the internal test, with AUROC of 0.83 in the SP2 dataset and an AUROC of 0.65 in the BES dataset for external test (Table S5, available at www.ophthalmologyscience.org).

Discussion

To our knowledge, this study represents one of the first to integrate DIR technique into DL models for prediction of refractive errors using retinal photographs. The DIR-integrated models consistently demonstrated superior MAE and R^2 performance compared with their baseline models (Tables 3 and 4). Deep imbalanced regression's improved accuracy in handling imbalanced datasets suggests greater suitability for real-world evaluation where data imbalance is common. These findings highlight the potential utility of combining DL models with retinal imaging for opportunistic screening of refractive errors, particularly in settings where retinal cameras are already in use.

Our results are comparable to previous DL studies using retinal images to predict refractive error. Varadarajan et al (2018) reported a slightly lower MAE (spherical: 0.56D, SE: 0.56D–0.91D), while Zou et al (2022) reported an MAE of 0.50D to 0.63D for spherical power prediction. However, both studies demonstrated lower R^2 values, suggesting that despite achieving better MAE metrics, their models may not effectively capture underlying patterns, thereby limiting generalizability. Additionally, we also explored cylinder power prediction and observed suboptimal performance (Table S6, available at www.ophthalmologyscience.org). This is not unexpected as cylinder power is mainly attributed to cornea and lens and not retinal.[40] We conducted sensitivity analyses by removing eyes with age-related macular degeneration and DR (Table S7, available at www.ophthalmologyscience.org). Overall, the MAE

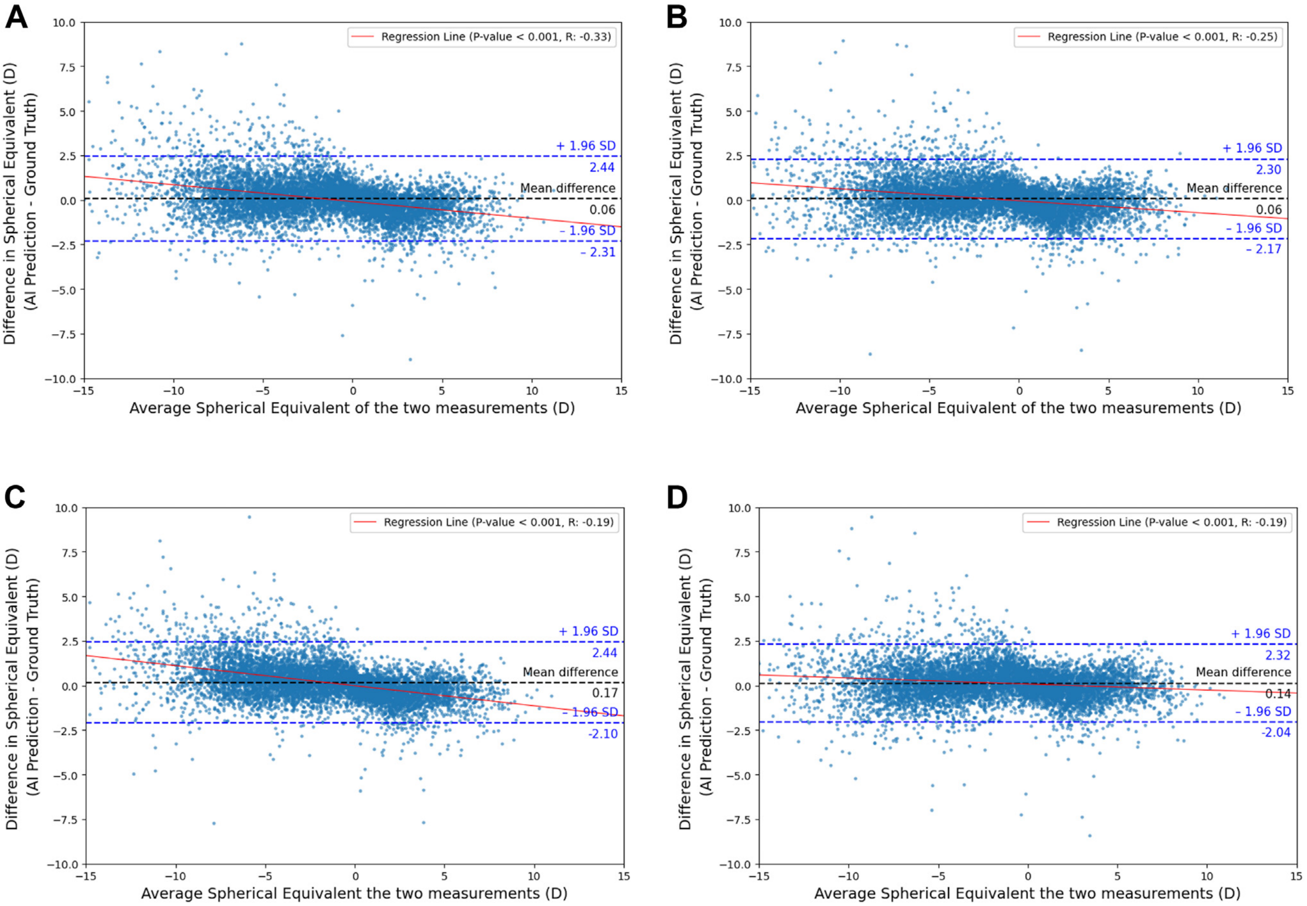

Figure 7. Bland−Altman plots of spherical equivalent in internal test: **A,** model: ResNet34 (MAE: 0.81D, R^2: 0.93); **B,** model: ResNet34 with DIR (MAE: 0.78D, R^2: 0.94); **C,** model: SwinV2 (MAE: 0.78D, R^2: 0.93); **D,** model: SwinV2 with DIR (MAE: 0.75D, R^2: 0.94). Red line represents the least squares regression line. DIR = deep imbalanced regression; MAE = mean absolute error; SD = standard deviation.

performance remained similar with the original analyses (Tables 3 and 4). Furthermore, models with DIR continued to consistently outperform those without DIR in these sensitivity analyses (Table S7, available at www.ophthalmologyscience.org). We also performed subgroup analyses on the higher hyperopic (+5D or worse) and higher myopic (−10D or worse) groups and observed that DIR-integrated models yielded lower MAE compared with non-DIR models across both tasks (Table S8, available at www.ophthalmologyscience.org), indicating that DIR method may be applied to higher refractive error range as well. However, it is important to note that the MAE values in these subgroups were higher than those observed in the main analyses (Tables 3 and 4).

We also observed that the BES dataset consistently demonstrated lower MAE than internal dataset for both spherical power and SE predictions (Tables 3 and 4). As shown in the Bland−Altman plots for internal test sets (Figs 6 and 7), our models consistently perform better on lower refractive error range (i.e., −5.00D to +5.00D). Coincidentally, a large portion of the BES dataset (98%) falls within range as well (Figs S9 and S11, available at www.ophthalmologyscience.org). This may have attributed to the lower MAE observed in BES.

Our study is unique in its application of the DIR technique to ResNet34 and SwinV2 to address data imbalances in refractive error prediction. Traditional data imbalance handling techniques such as Synthetic Minority Over-sampling TEchnique-based regression methods (i.e., Synthetic Minority Over-sampling TEchnique for Regression[41] and Synthetic Minority Over-sampling TEchnique for Regression with Gaussian Noise[42]) address data imbalance by generating synthetic samples for predefined rare regions or ranges. However, these traditional methods may have limitations when applied to high-dimensional data and computer vision tasks. In contrast, the DIR approach, which incorporates LDS and feature distribution smoothing, has proven more effective in handling data imbalance in computer vision task, yielding better overall performance.[12] We applied the DIR approach to our models, resulting in a lower MAE and higher R^2 for predicting spherical power and SE, and with vision transformer models generally achieving superior performance than convolutional neural network (Tables 3 and 4). The observed reductions in MAE across all DIR models, compared with their non-DIR counterparts, further underscore the robustness of the DIR method across various model architectures. Secondly, we trained our model on a multiethnic cohort from the SEED and UKBB datasets and validated it externally with the SP2 and BES datasets, a step omitted in previous research.[13–18] The design which combined the SEED and UKBB datasets, was motivated by the objective of developing a more diverse and potentially more generalizable model. Thirdly, we split our training, validation, and test sets by a balanced approach for regression tasks, ensuring each set included an appropriate number of images from each range of refractive error to fairly represent the full spectrum. These approaches enhanced the robustness and credibility of the DL algorithm's performance.

Non-DL methods, such as autorefraction and subjective refraction are common methods to measure refractive error. However, autorefraction requires the acquisition of an additional instrument, whereas subjective refraction is more time consuming.[43] Although retinal photo-based DL models may not match the accuracy of autorefraction, they could be more practical in screening settings where retinal cameras are already in use (e.g., DR screening programs), offering opportunistic screening for refractive errors. Nonetheless, their integration into routine clinical practice still requires careful consideration.

Our study had several limitations. Firstly, while incorporating DIR into both ResNet34 and SwinV2 model enhanced predictive accuracy, certain challenges remain. All models tend to slightly underpredict refractive error in the hyperopic range and overpredict in the myopic range (Figs 6 and 7, and Figs S8–S11, available at www.ophthalmologyscience.org). Scatter points are more tightly clustered in the low refractive error range, suggesting better performance in this range which is likely due to the higher proportion of low refractive errors in the training data. Despite this, DIR-integrated models consistently outperformed their baseline counterparts (Tables 3 and 4).

Next, the use of different refractometry methods and retinal image capture conditions may introduce variability, particularly in the model's performance during external testing. However, this variability also allows us to evaluate the model's robustness and generalizability across different clinical settings. By utilizing diverse datasets, we can better assess the model's performance where conditions and equipment may vary, thereby offering a more comprehensive assessment of its clinical utility.

The DIR-integrated DL models have shown potential in addressing data imbalances in predicting refractive error from retinal photographs. These findings highlight the potential utility of combining DL models with retinal imaging for opportunistic screening of refractive errors, particularly in settings where retinal cameras are already in use.

Footnotes and Disclosures

Originally received: June 11, 2024.
Final revision: November 17, 2024.
Accepted: November 18, 2024.
Available online: November 28, 2024.

[1] Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore.

[2] Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore.

[3] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore.

[4] Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore.

[5] Beijing Ophthalmology and Visual Science Key Lab, Beijing Institute of Ophthalmology, Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China.

[6] Institute of Molecular and Clinical Ophthalmology, Basel, Switzerland.

[7] Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany.

[8] Ophthalmology and Visual Sciences (Eye ACP), Duke-NUS Medical School, Singapore, Singapore.

*S.M.E.Y and X.L. contribute equally as first author.

[†]C.Y.C and Y.C.T. contribute equally as last author.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

This work was supported by the Agency for Science, Technology and Research (A*STAR) under its RIE2020 Health and Biomedical Sciences (HBMS) Industry Alignment Fund Pre-Positioning (IAF-PP, Grant Number: H20c6a0031). Y.C.T. is supported by the National Medical Research Council of Singapore (NMRC/MOH/ HCSAINV21nov-0001).

HUMAN SUBJECTS: Human subjects were included in this study. Written informed consent was obtained from all participants. Each study complied with the principles of the Declaration of Helsinki and received approval from the respective local ethical committees. SEED and SP2 datasets were approved by SingHealth Centralised Instituitional Review Board (CIRB). UKBB dataset was approved by North West Multi-centre Research Ethics Committee (MREC). BES dataset was approved by The Medical Ethics Committee of the Beijing Tongren Hospital. Additionally, data usage permission was granted by the principal investigator of each study.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Yew, Lei, Chen, Goh, Pushpanathan, Xue, Xu, Liu, Cheng, Tham

Data collection: Yew, Lei, Chen, Goh, Wang, Jonas, Sabanayagam, Tham

Analysis and interpretation: Yew, Lei, Pushpanathan, Xue, Sabanayagam, Koh, Xu, Liu, Cheng, Tham

Obtained funding: N/A

Overall responsibility: Yew, Lei, Chen, Goh, Pushpanathan, Xue, Wang, Jonas, Sabanayagam, Koh, Xu, Liu, Cheng, Tham

Manuscript no. XOPS-D-24-00185.

Abbreviations and Acronyms:

AUROC = area under the receiver operating characteristics curve; **BES** = Beijing Eye Study; **DIR** = Deep Imbalanced Regression; **DL** = deep learning; **DR** = diabetic retinopathy; **LOA** = limits of agreement; **MAE** = mean absolute error; **R2** = coefficient of determination; **SE** = spherical equivalent; **SEED** = Singapore Epidemiology of Eye Diseases; **SP2** = Singapore Prospective Study; **UKBB** = United Kingdom Biobank; **VI** = visual impairment.

Keywords:

Deep learning, Imbalanced regression, Refractive error, Retinal photos.

Correspondence:

Dr. Yih-Chung Tham, PhD, Yong Loo Lin School of Medicine, National University of Singapore, Level 13, MD1 Tahir Foundation Building, 12 Science Drive 2, Singapore 117549. E-mail: thamyc@nus.edu.sg.

References

1. Organization WH. *World Report on Vision*. Geneva: World Health Organization; 2019.
2. Honavar SG. The burden of uncorrected refractive error. *Indian J Ophthalmol*. 2019;67:577−578.
3. Organization WH. *Report of the 2030 Targets on Effective Coverage of Eye Care*. Geneva: World Health Organization; 2022.
4. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA Network*. 2016;316:2402−2410.
5. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye Diseases using retinal images from multiethnic populations with diabetes. *JAMA Network*. 2017;318:2211−2223.
6. Grassmann F, Mengelkamp J, Brandl C, et al. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*. 2018;125:1410−1420.
7. Keel S, Li Z, Scheetz J, et al. Development and validation of a deep-learning algorithm for the detection of neovascular age-related macular degeneration from colour fundus photographs. *Clin Exp Ophthalmol*. 2019;47:1009−1018.
8. Peng Y, Dharssi S, Chen Q, et al. DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*. 2019;126:565−575.
9. Tham Y, Goh JHL, Anees A, et al. Detecting visually significant cataract using retinal photograph-based deep learning. *Nat Aging*. 2022;2:264−271.
10. Tham YC, Anees A, Zhang L, et al. Referral for disease-related visual impairment using retinal photograph-based deep learning: a proof-of-concept, model development study. *Lancet Digit Health*. 2021;3:e29−e40.
11. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol*. 2019;137:1353−1360.
12. Yang Y, Zha K, Chen Y, et al. Delving into deep imbalanced regression. In: Marina M, Tong Z, eds. *Proceedings of the 38th International Conference on Machine Learning*. Cambridge, Massachusetts: Proceedings of Machine Learning Research, PMLR; 2021:11842−118451.
13. Chun J, Kim Y, Shin KY, et al. Deep learning-based prediction of refractive error using photorefraction images captured by a smartphone: model development and validation study. *JMIR Med Inform*. 2020;8:e16225.
14. Shi Z, Wang T, Huang Z, et al. A method for the automatic detection of myopia in Optos fundus images based on deep learning. *Int J Numer Method Biomed Eng*. 2021;37:e3460.
15. Varadarajan AV, Poplin R, Blumer K, et al. Deep learning for predicting refractive error from retinal fundus images. *Invest Ophthalmol Vis Sci*. 2018;59:2861−2868.
16. Xu D, Ding S, Zheng T, et al. Deep learning for predicting refractive error from multiple photorefraction images. *Biomed Eng Online*. 2022;21:55.
17. Yoo TK, Ryu IH, Kim JK, Lee IS. Deep learning for predicting uncorrected refractive error using posterior segment optical coherence tomography images. *Eye (Lond)*. 2022;36:1959−1965.
18. Zou H, Shi S, Yang X, et al. Identification of ocular refraction based on deep learning algorithm as a novel retinoscopy method. *Biomed Eng Online*. 2022;21:87.
19. Tan TE, Anees A, Chen C, et al. Retinal photograph-based deep learning algorithms for myopia and a blockchain platform to facilitate artificial intelligence medical research: a retrospective multicohort study. *Lancet Digit Health*. 2021;3:e317−e329.
20. Yang D, Li M, Li W, et al. Prediction of refractive error based on ultrawide field images with deep learning models in myopia patients. *Front Med (Lausanne)*. 2022;9:834281.
21. Yang Y, Li R, Lin D, et al. Automatic identification of myopia based on ocular appearance images using deep learning. *Ann Transl Med*. 2020;8:705.
22. Ghosh K, Bellinger C, Corizzo R, et al. The class imbalance problem in deep learning. *Mach Learn*. 2022;113:4845−4901.
23. Zhang Y, Kang B, Hooi B, et al. Deep long-tailed learning: a Survey2021. https://ui.adsabs.harvard.edu/abs/2021arXiv211004596Z. Accessed October 1, 2021.
24. de Hond AAH, Leeuwenberg AM, Hooft L, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med*. 2022;5:2.
25. He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, USA: Institute of Electrical and Electronics Engineers (IEEE); 2016:770−778.
26. Liu Z, Hu H, Lin Y, et al. *Swin Transformer v2: Scaling up Capacity and Resolution. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. New Orleans, LA: Institute of Electrical and Electronics Engineers (IEEE); 2022:12009−12019.
27. Majithia S, Tham YC, Chee ML, et al. Cohort profile: the Singapore Epidemiology of eye Diseases study (SEED). *Int J Epidemiol*. 2021;50:41−52.
28. Chua SYL, Thomas D, Allen N, et al. Cohort profile: design and methods in the eye and vision consortium of UK Biobank. *BMJ Open*. 2019;9:e025077.
29. Tan KHX, Tan LWL, Sim X, et al. Cohort profile: the Singapore multi-ethnic cohort (MEC) study. *Int J Epidemiol*. 2018;47:699−j.
30. Jonas JB, Xu L, Wang YX. The Beijing Eye study. *Acta Ophthalmol*. 2009;87:247−261.
31. Rim TH, Soh ZD, Tham YC, et al. Deep learning for automated sorting of retinal photographs. *Ophthalmol Retina*. 2020;4:793−800.
32. Cui Y, Jia M, Lin TY, et al. *Class-balanced loss based on effective number of samples. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. Long Beach, CA: Institute of Electrical and Electronics Engineers (IEEE); 2019:9268−9277.
33. Wong TY, Tham YC, Sabanayagam C, Cheng CY. Patterns and risk factor profiles of visual loss in a multiethnic asian population: the Singapore Epidemiology of eye Diseases study. *Am J Ophthalmol*. 2019;206:48−73.
34. Leone JF, Mitchell P, Morgan IG, et al. Use of visual acuity to screen for significant refractive errors in adolescents: is it reliable? *Arch Ophthalmol*. 2010;128:894−899.
35. McDonnell CE. *Refraction and Prescribing*. *Dispensing optics, March 2012*. England: Dispensing Optics Journal, Association of British Dispensing Opticians (ABDO); 2012.
36. Wolffsohn JS, Bhogal G, Shah S. Effect of uncorrected astigmatism on vision. *J Cataract Refract Surg*. 2011;37:454−460.

37. Black AA, Wood JM, Colorado LH, Collins MJ. The impact of uncorrected astigmatism on night driving performance. *Ophthalmic Physiol Opt*. 2019;39:350–357.
38. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.
39. Bland JM, Altman DG. Agreed statistics: measurement method comparison. *Anesthesiology*. 2012;116:182–185.
40. Gurnani B, Kaur K. *Astigmatism*. Treasure Island (FL): StatPearls; 2023.
41. Torgo L, Ribeiro RP, Pfahringer B, Branco P. SMOTE for regression. In: Correia L, Reis LP, Cascalho J, eds. *Progress in Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013:378–389, 2013 2013//.
42. Branco P, Torgo L, Ribeiro RP. SMOGN: a pre-processing approach for imbalanced regression. In: Paula Branco Luís T, Nuno M, eds. *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*. Porto, Portugal: Proceedings of Machine Learning Research, PMLR; 2017:36–50.
43. Lopez VR, Hernandez-Poyatos A, Dorronsoro C. The Direct Subjective Refraction: unsupervised measurements of the subjective refraction using defocus waves. *bioRxiv*. 2021;2021. https://doi.org/10.1101/2021.12.04.471123.