Communications Medicine

https://doi.org/10.1038/s43856-026-01573-y

Article in Press

High-accuracy retinal age prediction via fundus-based multitask learning reveals the effect of systemic disease

Received: 1 July 2025

Accepted: 23 March 2026

Cite this article as: Ninomiya, T., Hanyuda, A., Kiyota, N. *et al.* High-accuracy retinal age prediction via fundus-based multitask learning reveals the effect of systemic disease. *Commun Med* (2026). https://doi.org/10.1038/s43856-026-01573-y

Takahiro Ninomiya, Akiko Hanyuda, Naoki Kiyota, Parmanand Sharma, Yukun Zhou, Siegfried K. Wagner, Keita Suzuki, Takanari Nozaki, Takehiro Miya, Naoki Takahashi, Kazuko Omodaka, Noriko Himori, Yoichi Ichikawa, Pearse A. Keane & Toru Nakazawa

Article

High-accuracy retinal age prediction via fundus-based multitask learning reveals the effect of systemic disease

Takahiro Ninomiya[1,2], Akiko Hanyuda[1,3], Naoki Kiyota[1], Parmanand Sharma[1], Yukun Zhou[2], Siegfried K. Wagner[2,4], Keita Suzuki[5], Takanari Nozaki[5], Takehiro Miya[1], Naoki Takahashi[1], Kazuko Omodaka[1], Noriko Himori[1,6], Yoichi Ichikawa[5], Pearse A. Keane[2,4], Toru Nakazawa[1,7,8,9*]

Affiliations:

[1] Department of Ophthalmology, Tohoku University Graduate School of Medicine, Miyagi, Japan.

[2] UCL Institute of Ophthalmology, London, United Kingdom.

[3] Department of Ophthalmology, Keio University School of Medicine, Tokyo, Japan.

[4] NIHR Moorfields Biomedical Research Centre, London, United Kingdom.

[5] The Dai-ichi Life Insurance Company, Limited, Tokyo, Japan.

[6] Department of Aging Vision Healthcare, Tohoku University Graduate School of Biomedical Engineering, Tohoku University, Sendai, Japan.

[7] Department of Ophthalmic Imaging and Information Analytics, Tohoku University Graduate School of Medicine, Miyagi, Japan.

[8] Department of Retinal Disease Control, Tohoku University Graduate School of Medicine, Miyagi, Japan.

[9] Department of Advanced Ophthalmic Medicine, Tohoku University Graduate School of Medicine, Miyagi, Japan.

***Correspondence:** Toru Nakazawa, MD, PhD

Department of Ophthalmology, Tohoku University Graduate School of Medicine,

1-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8574, Japan.

Tel: +81 22 717 7294, Fax: +81 22 717 7298

E-mail: toru.nakazawa.e1@tohoku.ac.jp

Abstract

Background: Accurate estimation of the retinal age, defined as the age predicted from fundus photographs by a deep-learning model trained on chronological age, provides a non-invasive biomarker of biological ageing and disease risk.

Methods: In this study, we trained an ensemble multitask learning model that integrates fundus photographs with glycated haemoglobin using 50,595 quality-controlled fundus photographs from 27,214 disease-free adults and validated it on an independent set of 7,288 additional images from disease-free adults. Model performance was evaluated using mean absolute error. Prediction uncertainty was quantified by calculating the standard deviation across ensemble predictions for each eye, and eyes were stratified based on this standard deviation.

Results: Here we show that the model achieves mean absolute errors of 2.78 years in internal validation and 3.39 years and 8.63 years in two external cohorts comprising 135 and 4,992 eyes, respectively. Eyes with ensemble standard deviations below the median demonstrate improved age-prediction accuracy (mean absolute error: 2.46 years internally; 2.87 years in the primary external cohort). In a systemic disease cohort of 8,467 individuals, the retinal age gap (predicted minus chronological age) is significantly higher in participants with diabetes, cardiac disease, or stroke after adjustment for age and sex, indicating older-appearing retinas and supporting the biological relevance of retinal age.

Conclusions: Retinal age derived from a single fundus photograph provides a scalable and clinically deployed biomarker of biological ageing. This approach may enable opportunistic screening for cardiometabolic and other ageing-related diseases in other routine ocular imaging workflows.

Plain Language Summary

Retinal photographs are images of the inner, back surface of the eye and are routinely taken in eye clinics. Aging alters the retinal appearance. We built a computational model that estimates the age of the retina from a single retinal photo. The system was trained on 50,595 images from 27,214 disease-free adults. Predictions of age were most accurate when high-quality images were used and remained accurate when tested on additional images. A large gap between a person's actual age and the estimated retinal age was found more in people with diabetes, heart disease or stroke. This tool could therefore be helpful to assess cardiometabolic health during routine eye appointments.

Introduction

Artificial intelligence (AI) has transformed the extraction of clinically relevant information from medical images, particularly in ophthalmology[1,2]. Retinal imaging techniques such as fundus photography and optical coherence tomography (OCT) provide a non-invasive view of ocular and systemic circulation, as highlighted in oculomics frameworks that link ocular findings to systemic disease[3–5]. Deep-learning systems are able to detect major eye diseases, including diabetic retinopathy, glaucoma, and age-related macular degeneration, with high accuracy in large multiethnic cohorts, as illustrated by models for diabetic retinopathy screening and glaucomatous optic neuropathy[6–10].

In 2018, Poplin et al. reported that a convolutional neural network could estimate chronological age from fundus images with a mean absolute error of 3.26 years, focusing attention on age prediction from the retina[11]. The difference between retinal-predicted and chronological age—often termed the retinal age gap (RAG)—has been associated with cardiovascular disorders, including stroke and incident cardiovascular disease mediated by increased arterial stiffness[11–16]. Large cohort studies have further demonstrated that retinal ageing markers, including RAG and related indices, predict a wide range of outcomes such as chronic obstructive pulmonary disease, all-cause morbidity and mortality, kidney failure and Parkinson's disease, supporting their potential role as general indicators of systemic ageing and vulnerability[13,14,17–19].

A recent scoping review of retinal age estimation studies published between 2022 and 2023 reported mean absolute errors (MAEs) ranging from 3.30 to 3.97 years[20], while complementary work has summarised ocular ageing biomarkers and cross-population retinal ageing models[21,22]. Taken together, these reports suggest that current model architectures may be approaching a performance ceiling. Moreover, many retinal age studies provide limited information on network design or rely heavily on pre-existing models, which can hinder interpretability and translation into diverse clinical environments[13,23]. Even when promising results are achieved, the performance of age-prediction models

often degrades when applied to images acquired with different cameras or in populations that differ from the training cohort in image quality or demographic characteristics, underscoring the need for robust, generalisable models[24].

To address these gaps, we aimed to develop a robust retinal age model trained on fundus photographs from a large health-check cohort, using auxiliary blood-test information only during training. We focused on simple architectural choices: a two-output multitask learning (MTL) framework in which age and a systemic biomarker are predicted together, and an ensemble of fold-specific models to stabilise predictions. We restricted the MTL architecture to two regression outputs, age, and exactly one systemic biomarker to maintain architectural simplicity, facilitate interpretation of the effect of adding a single auxiliary prediction task, and reduce the risk of negative transfer from multiple heterogeneous tasks. Among the candidate auxiliary outputs (glycated haemoglobin [HbA1c], systolic blood pressure, and low-density lipoprotein cholesterol), we selected HbA1c because HbA1c is a clinically established and temporally stable marker of metabolic status. The deployed model requires only a single fundus photograph at inference as systemic measurements are used solely as auxiliary training signals, enabling an image-only workflow compatible with routine ocular imaging and scalable screening.

We show that this approach achieves low error in internal validation and remains accurate in external cohorts, and that disagreement across the ensemble provides a practical confidence score that identifies images with higher prediction accuracy. We further show that a larger retinal age gap is associated with diabetes, cardiac disease, and stroke in a systemic disease cohort, supporting retinal age as an imaging biomarker of systemic ageing.

Methods

Study population and data collection

This study analysed data from 64,772 individuals who participated in health screening at the Miyagi Center for Lifestyle Disease Prevention between 2020 and 2022, yielding 157,695 eye records. We applied inclusion and exclusion criteria to construct a dataset focusing on the relationship between retinal imaging characteristics and systemic health indicators. We excluded 14,183 individuals with missing key variables (age, sex, HbA1c, LDL cholesterol, blood pressure, or smoking history), eliminating 38,831 eye records.

To establish a longitudinal cohort for future analyses, we identified 10,356 individuals (35,998 eye records) with a continuous 3-year follow-up period. This cohort will be used in subsequent work to examine how the RAG (predicted minus chronological age) evolves over time and how temporal changes relate to the progression of systemic disease.

The remaining cohort was further refined by excluding individuals with a history of cerebrovascular or cardiovascular disease and those currently taking antihypertensive, lipid-lowering, or hypoglycaemic agents. We also excluded participants with blood pressure >140/90 mmHg, body mass index <18.5 or >25 kg/m², fasting plasma glucose >126 mg/dL, or HbA1c >6.5%. Because the training and internal-validation sets were intended to model and evaluate age-related changes in ostensibly healthy eyes, the same exclusion criteria were applied to these subsets.

After these steps, 27,214 individuals (50,595 eye records) were assigned to the training cohort and a non-overlapping set of 7,288 individuals to the internal-validation cohort. These cohorts were used to develop and internally validate the retinal age model.

To investigate associations between RAG and systemic disease, we next derived a systemic disease assessment set from the same 3-year follow-up cohort. After discarding poor-quality images and retaining only the first examination for each participant, 8,467 individuals met the inclusion criteria. When both eyes were available, one eye was selected to avoid within-participant correlation: if image quality differed, the higher-quality eye was chosen; when quality was similar, one eye was selected at

random. The resulting 8,467 eye records constituted the systemic disease assessment set used to analyse relationships between RAG and systemic conditions. Fig. 1 summarises the inclusion and exclusion processes.

In addition, to validate the robustness of our AI model in an external setting, we used two independent health-screening cohorts. The first external cohort (External Test 1) comprised 135 individuals who underwent health screening at Sendai Open Hospital and had at least one gradable fundus photograph. Participants were selected using clinical inclusion and exclusion criteria that matched those for the training and internal-validation cohorts as closely as possible; however, individuals with missing HbA1c values were retained because HbA1c was used only as an auxiliary target during training and is not required for model inference. When both eyes were available, one eye per participant was selected using the same rule as in the systemic disease assessment set. The second external cohort (External Test 2) was derived from the AlzEye 2018 study[25] and is described in detail in the “External validation datasets” subsection below.

Fundus image quality for the present study was reviewed by a board-certified ophthalmologist who routinely interprets fundus photographs acquired during health screening as part of daily clinical practice. Images with incomplete visualisation of the optic disc or macula, severe motion blur, media opacity, or large artefacts obscuring the major retinal vessels were classified as poor quality and excluded before model development.

The study procedures were approved by the Institutional Review Board of Tohoku University Graduate School of Medicine (No. 2021-1-265) and conducted in accordance with the tenets of the Declaration of Helsinki. The development and internal validation dataset was analysed as part of a retrospective study approved by the Institutional Review Board of Tohoku University Graduate School of Medicine (No. 2021-1-265). Due to the retrospective nature of the study, the requirement for informed consent was waived by the Institutional Review Board, and an opt-out approach was

implemented in accordance with institutional guidelines. The data were accessed by the research team under institutional approval. The external-validation dataset from Sendai Open Hospital (External Test 1) consisted of retinal images obtained at Sendai Open Hospital. The study was conducted under ethical approval from the Institutional Review Board of Sendai Open Hospital (Approval No. 2019-0038) and the Institutional Review Board of Tohoku University Graduate School of Medicine (Approval No. 2025-1-680). For this dataset, written informed consent was obtained directly from participants at the time of study enrolment, and only individuals who provided consent were included. The data were accessed through an approved inter-institutional collaboration between Sendai Open Hospital and Tohoku University. The External Test 2 dataset was derived from the AlzEye 2018 study and was analysed under the AlzEye governance framework and the ethics approvals described for the AlzEye project (London-Central Research Ethics Committee approval 18/LO/1163 and Confidential Advisory Group approval for Section 251 support 18/CAG/0111, waiving the need for consent, with final approval by the National Health Service Health Research Authority). AlzEye is a retrospective cohort based on routinely collected NHS ophthalmic and hospital admissions data. All methods followed relevant guidelines and regulations.

External validation datasets

For the second external-validation dataset (External Test 2), we incorporated a subset of the AlzEye 2018 cohort[25], representing a health-screening population. To approximate as closely as possible the inclusion and exclusion criteria used for the training and internal-validation cohorts within the constraints of the available variables, we restricted this dataset to participants of Asian ethnicity and excluded individuals with a history of hypertension, cardiac disease, or cerebrovascular disease according to the available clinical records. Because only diagnosis- or history-based information was available for systemic diseases and laboratory measurements such as HbA1c were not accessible, we

used these self-reported or coded disease histories (for example, a history of diabetes) as surrogates for corresponding laboratory-based criteria in the development cohorts. This restriction was chosen to reduce ethnic domain shift relative to the Japanese development cohorts while keeping the analysis computationally tractable for the present study.

Because fundus imaging records in AlzEye 2018 often contained multiple examinations per participant, we selected one eye per person based on a predefined rule: the right eye was preferentially chosen, and when multiple visits were available, the earliest examination date was selected. Owing to institutional data-governance restrictions, raw-level clinical data from this institution could not be transferred across sites. As a result, baseline comparisons with External Test 2 were based on summary statistics (means, SDs, and sample sizes), and model-performance summaries were computed and shared in aggregated form only.

Measurement of clinical variables

The datasets for the training, internal-validation, and external cohorts included clinical data collected during health screening, comprising interview-based medical history, medication use, routine blood tests (including blood glucose and lipid profiles), and physical measurements such as height, weight, and blood pressure. These variables were used to characterise participants' systemic health and relate systemic factors to retinal imaging findings.

Retinal photographs for the training and internal-validation cohorts were captured using a non-mydriatic fundus camera (CR-2 AF; Canon Inc., Tokyo, Japan). External Test 1 and External Test 2 used images acquired with a different device (DRI-OCT Triton; Topcon Corporation, Tokyo, Japan), allowing us to assess model performance across imaging platforms.

For the systemic disease analyses, medication variables were obtained from self-reported questionnaires. For each drug class (antihypertensive, hypoglycaemic, or lipid-lowering agents), "Yes"

indicated current use of the corresponding medication, whereas “No” indicated no such reported use; therefore, the non-medication group may include both disease-free individuals and those with untreated or undiagnosed conditions.

Machine learning methods

To identify the optimal model, we first performed model-comparison experiments on a reduced subset of the training data. Approximately 5% of eye records were randomly sampled from the training cohort after application of all inclusion and exclusion criteria, yielding a reduced training set whose demographic and clinical characteristics closely reflected those of the full development cohorts (Table 1). Using this reduced dataset, we trained the candidate single-task learning (STL) and MTL configurations listed in Table 2 with five-fold cross-validation. In each fold, models were trained on four fifths of the subset and evaluated on the independent internal-validation cohort (n = 7,288), and the MAE, MSE, and MAPE for age prediction were recorded. For MTL models, the performance metrics reported in Table 2 refer to the age-prediction output; the auxiliary systemic-variable output (for example, HbA1c) was included as a regularising task and was not used as a primary model-selection criterion.

On the basis of these comparisons, we selected the MTL configuration that used fundus images to jointly predict age and HbA1c as the final architecture, because it achieved the lowest ensemble MAE and showed a statistically significant improvement over the baseline STL model (Table 2). This selected architecture was then trained on the full training cohort using five-fold cross-validation at the participant level to generate five fold-specific models. Each fold-specific model was evaluated on the same independent internal-validation cohort (n = 7,288), which was not used during training. The final ensemble model was obtained by computing the arithmetic mean of the predictions from these five models on each eye, thereby enhancing the stability and reducing the variance of age estimates. For each

eye, we also calculated the SD of the five fold-specific age predictions as an internal proxy for prediction uncertainty, with larger SD indicating greater disagreement among models. This SD was used to stratify eyes into Low-SD and High-SD subgroups at the cohort-specific median in subsequent accuracy and systemic-disease analyses. The ensemble was then compared head-to-head with the publicly available JOIR retinal age model and with baseline and fine-tuned RETFound-DINOv2 models in the external cohorts (Tables 3 and 4)[26-28].

In this study, we implemented both STL and MTL models. Fig. 2 illustrates the architecture of the models. The STL model (Fig. 2a) processes retinal images, producing a single regression output, age. This model follows a conventional convolutional neural network (CNN) structure with fully connected layers (fc) for age prediction. For STL configurations that used additional non-image inputs ('Fundus + HbA1c' or 'Fundus + Sex'), these covariates were standardised and concatenated with the convolutional feature vector before the fully connected regression head, allowing the network to jointly exploit retinal and systemic information.

In contrast, the MTL model (Fig. 2b) attaches two parallel fc heads to the same convolutional backbone so that a single fundus image simultaneously yields two continuous outputs—age and a second biometric parameter (e.g. HbA1c). Because both outputs are optimised under a joint loss, the shared convolutional weights are updated by information from both tasks, yielding feature representations that differ from those learned in the STL setting.

For the convolutional architecture ('CNN' in Fig. 2), we based our model on the LWBNA_Unet (Lightweight Bottleneck Narrowing with Attention in Unet) structure reported by Sharma et al. in 2022 and enhanced it by incorporating residual blocks into the convolution layers to improve feature extraction capabilities[29]. For model interpretation, we computed Grad-CAM heat maps[30] from the last convolutional block to visualise image regions that contributed most to each prediction (Fig. 3). For each image, the Grad-CAM values were min–max normalised, mapped to a

colour scale in which warmer colours indicate relatively higher contribution, and overlaid semi-transparently on the original fundus photograph; these maps were used as qualitative saliency visualisations rather than quantitative estimates of feature importance. For each image, Grad-CAM activations were normalised to the [0, 1] range and rendered with a fixed heatmap such that higher-intensity regions indicate greater relative contribution to the predicted age; the same normalisation and colour scale were used for all fold-specific models and for the ensemble to allow qualitative comparison of saliency patterns.

For comparison with a retinal foundation model, we additionally evaluated RETFound-DINOv2, a self-supervised transformer foundation model for retinal images that builds on the RETFound architecture and has been trained on large multi-site retinal cohorts including the AlzEye 2018 dataset[26,27]. In this study, we defined the RETFound-DINOv2 baseline as an age-prediction model that used a RETFound-DINOv2 backbone trained on multi-site retinal datasets including AlzEye 2018, with the original classification head replaced by a regression head for age; a fine-tuned RETFound-DINOv2 model was then obtained by further fine-tuning the regression head and upper backbone layers of this baseline model on our training cohort, following the general strategy described by Zhou et al.[26]. Early stopping in this second stage was based on performance in the internal-validation cohort. The underlying vision transformer design of RETFound-DINOv2 is related to the general-purpose DINOv2 framework for self-supervised pre-training of large vision models on natural images[31].

The models were optimised using the Lion optimiser, with a learning rate set specifically for each model[32]. The loss function combined a log-cosh data-fitting term with elastic-net regularisation, designed to robustly handle the variability and outliers typical of medical image data[33, 34]. For the MTL model, the data-fitting component of the loss was defined as

$$L_{MTL} = \lambda_{age} L_{age} + \lambda_{parameter\ 2} L_{parameter\ 2}$$

where L_{age} and $L_{parameter\ 2}$ represent the log-cosh losses for age and HbA1c prediction tasks,

respectively, with λ_{age} and $\lambda_{parameter\ 2}$ balancing the importance of each task.

Early stopping was implemented after 10 epochs to prevent overfitting and ensure an efficient training time. This strategy is crucial for maintaining the generalisability of the model across different datasets. The computational work for these machine learning analyses was conducted using Python (version 3.11.7) and PyTorch (version 2.2.2) with hardware support from a 2 × Intel Xeon Gold 6648Y CPU and a 4 × NVIDIA A6000 Ada GPU, ensuring robust data processing capabilities.

Statistics and reproducibility

All statistical analyses were performed using Python (version 3.11.7; scikit-learn version 1.2.2) and R (version 4.3.2). Continuous variables are summarised as means ± SDs or medians with interquartile ranges, and categorical variables as counts with percentages.

For baseline characteristics (Table 1), group comparisons between the training, internal-validation, and External Test 1 cohorts (A–C) used two-tailed Mann–Whitney *U* tests for continuous variables and χ^2 tests for categorical variables when individual-level data were available. For External Test 2 (cohort D), only summary statistics were available; *P*-values for comparisons with the training cohort were therefore estimated from means, SDs, and sample sizes using Welch's two-sample *t*-test for continuous variables and the χ^2 test for categorical variables. To quantify between-cohort differences independent of the sample size, we calculated absolute standardised mean differences (|SMD|) for all pairwise cohort comparisons and, for completeness, report the corresponding two-sided *P*-values for each cohort pair (Supplementary Table 1).

To compare single-task and multitask model configurations in the initial model-selection stage (Table 2), we performed five-fold cross-validation within the reduced training cohort used for model-comparison experiments, generating five fold-specific models for each candidate configuration. These fold-specific models were then evaluated on the same independent internal-validation cohort, which had

been held out prior to model development. For each candidate model, the absolute error was calculated for each eye in the internal-validation cohort, and the difference in absolute error relative to the baseline STL model (fundus-only age prediction) was determined. Two-sided Wilcoxon signed-rank tests on these per-eye differences were used to obtain the P-values reported in Table 2. The 5-fold MAE, MSE, and MAPE values in Table 2 are presented as mean ± SD across the five fold-specific models, and the ensemble values were obtained by arithmetic averaging of the five model predictions on the same internal-validation eyes.

Model-performance metrics (MAE, MSE, and MAPE) for the internal-validation and external-validation analyses (Tables 3–5) are reported as mean ± SD calculated directly from the distribution of prediction errors for all eyes. Pairwise comparisons of model performance on the same eyes (for example, ensemble versus JOIR or RETFound-DINOv2) used two-sided Wilcoxon signed-rank tests.

For the systemic disease analysis (Table 6), participants who were taking antihypertensive, hypoglycaemic, or lipid-lowering medication or who reported cardiovascular disease, stroke, or chronic kidney disease were matched 1:1 to controls using nearest-neighbour propensity-score matching based on age and sex. RAGs between matched groups were compared with the Mann–Whitney *U* test. A two-sided $P < 0.05$ was considered statistically significant.

The code used for the analyses is publicly available (see Code Availability) and archived with a DOI for reproducibility[35].

Results

Participant characteristics of each dataset

The characteristics of the participants (Table 1) revealed considerable variations across the four datasets: the training cohort (A), internal-validation cohort (B), and two external-validation cohorts (C

and D). The training and internal-validation cohorts comprised 27,214 (50,595 eyes) and 7,288 (7,288 eyes) participants, respectively. External Test 1 (C), drawn from Sendai Open Hospital, contained 135 participants (135 eyes), whereas External Test 2 (D), derived from the AlzEye 2018 study[25], comprised 4,992 participants (4,992 eyes). Age distributions were broadly similar between the training and External Test 1 cohorts; the participants in the internal-validation cohort were slightly younger, whereas those in External Test 2 were somewhat older. External Test 1 had a markedly higher proportion of men (83.7%) than both the training (49.2%) and internal-validation (55.3%) cohorts.

Blood pressure and high-density lipoprotein (HDL) cholesterol levels were generally lower in External Test 1 than in the training and internal-validation cohorts, whereas low-density lipoprotein (LDL) cholesterol and HbA1c levels were broadly comparable across these three cohorts. External Test 2 demonstrated a higher mean age and modest differences in the blood pressure and lipid profiles relative to the development cohorts. As summarised by the P-values and absolute standardised mean differences in Table 1 and Supplementary Table 1, these findings indicate non-trivial clinical heterogeneity between the internal-validation and external cohorts, particularly for sex distribution, blood pressure, and HDL cholesterol. External Test 2 (AlzEye 2018) consisted of an Asian or Asian British subset filtered to match the clinical inclusion and exclusion criteria used in our training dataset: participants with hypertension, cardiac disease, or cerebrovascular disease were excluded according to available clinical records. One eye per participant was selected using a right-eye-priority rule, and the earliest examination date was used when multiple fundus examinations were available.

Performance of single-task and multitask models

Table 2 presents a comparison of age-prediction accuracy across various STL and MTL models; all metrics are computed for the age output in each configuration. Using the reduced training subset (approximately 5% of the eye records randomly sampled from the training cohort; see Methods),

we trained candidate STL and MTL configurations by five-fold cross-validation and compared them on the independent internal-validation cohort. The baseline STL model that used only fundus images as input achieved a five-fold MAE of 5.47 ± 0.50 years and an ensemble MAE of 5.03 years. Adding HbA1c or sex as additional input variables to the STL architecture did not improve performance; instead, ensemble MAEs increased slightly (5.19 and 5.15 years, respectively), with these degradations being statistically significant relative to the baseline model (ΔMAE +0.15 and +0.11 years; $P = 3.21 \times 10^{-5}$ and $P = 4.12 \times 10^{-4}$, respectively).

The MTL models demonstrated more diverse behaviour based on the combination of auxiliary outputs. The configuration that used fundus images and sex to jointly predict age and HbA1c yielded performance comparable to the baseline STL model (ensemble MAE, 5.05 years; ΔMAE, +0.02 years; $P = 0.284$). In contrast, the MTL model that used fundus images to predict both age and HbA1c achieved the lowest error among all candidates, with a five-fold MAE of 5.31 ± 0.33 years and an ensemble MAE of 4.86 years (ΔMAE −0.18 years; $P = 1.27 \times 10^{-4}$ vs. baseline). MTL variants predicting age together with systolic blood pressure or LDL cholesterol performed substantially worse (ensemble MAEs, 5.40 and 5.46 years; $P = 5.68 \times 10^{-6}$ and $P = 2.41 \times 10^{-6}$ vs. baseline, respectively). Based on these results, the MTL configuration with fundus-only input and dual outputs of age and HbA1c was selected as the final architecture for subsequent cross-validation, external-validation, and clinical-association analyses.

Cross-validation and ensemble model performance

Table 5 presents the results of five-fold cross-validation of the most effective MTL model, which used only fundus images to predict age and HbA1c levels. When evaluated on the internal-validation cohort, the MAE of the five fold-specific models ranged from 2.97 ± 2.56 years in Fold 5 to 3.12 ± 2.61 years in Fold 2, indicating consistent performance across folds. The ensemble

model, which aggregates predictions from all five folds, exhibited the best performance, with an MAE of 2.78 ± 2.39 years, mean squared error (MSE) of 13.42 ± 27.47 years², and mean absolute percentage error (MAPE) of 5.88 ± 5.36%.

Fig. 4 presents a detailed comparison of the actual and predicted ages from the ensemble model. The Bland–Altman plot (Fig. 4a) highlights that the differences between predicted and actual ages remain within acceptable limits, showing a bias of −0.05 years (± 7.18; 95% limits of agreement, −7.23 to 7.13 years). The scatterplot (Fig. 4b) shows a strong correlation (Pearson correlation coefficient $r = 0.97$, $P < 1 \times 10^{-300}$) between the actual and predicted ages, with the regression line closely following the ideal line of equality.

Fig. 3 presents the Gradient-weighted Class Activation Mapping (Grad-CAM) outputs[26] and age predictions for the five individual fold models and the ensemble model. The top row displays Grad-CAM maps that highlight the retinal regions each model attended to during prediction. In the bottom row, these heatmaps are super-imposed on the original fundus photographs, and the predicted age for each model is shown beneath each image. Across folds, these Grad-CAM maps consistently emphasised the optic disc, macula, and major temporal vascular arcades as key regions informing the age estimate, whereas the ensemble heatmap further concentrated attention on these structures while de-emphasising peripheral artefacts such as image borders.

Stratification by prediction standard deviation

Fig. 5 shows the stratification of the age-prediction accuracy based on the standard deviation (SD) of the predicted ages from the ensemble model. Because this SD reflects the degree of disagreement among the five fold-specific models and thus serves as an internal measure of prediction uncertainty, we divided eyes into Low-SD (higher confidence; SD below the cohort median of 1.71 years) and High-SD (lower confidence; SD at or above the median) subgroups. The scatterplot shows

the relationship between the absolute errors in age prediction and the predicted age SD, with lower SDs generally corresponding to smaller prediction errors. The box plot indicates a statistically significant difference ($P = 5.95 \times 10^{-22}$) between the Low-SD and High-SD groups in terms of prediction accuracy, with the Low-SD group showing the lowest error rates.

External validation and clinical associations

Table 3 presents a comparative analysis of the predictive performances in External Test 1 (Sendai Open Hospital; n = 135 eyes). Across all eyes, the MTL ensemble achieved an MAE of 3.39 ± 2.74 years, outperforming the Japan Ocular Imaging Registry (JOIR) retinal age model (5.37 ± 3.84 years; $P = 3.17 \times 10^{-7}$), fine-tuned RETFound-DINOv2 model (4.45 ± 3.15 years; $P = 0.002$), and RETFound-DINOv2 baseline (4.65 ± 3.31 years; $P = 1.04 \times 10^{-4}$). Similar patterns were observed for MSE and MAPE. In the Low-SD subgroup, the ensemble model showed the largest performance gain relative to all benchmark models, whereas in the High-SD subgroup, its MAE remained numerically lower than those of the benchmark models.

To further assess external generalisability, we applied the ensemble and JOIR models to External Test 2, an independent Asian subset of the AlzEye 2018 cohort[25] (4,992 eyes; Table 4). As expected, absolute errors were larger in this population-based cohort (overall MAE, 8.63 ± 6.33 years for the ensemble); however, the ensemble still significantly outperformed the JOIR model for MAE, MSE, and MAPE overall (all $P = 0.003$) and in the Low-SD subgroup (all $P = 0.008$). In the High-SD subgroup, differences between the two models were smaller and did not reach statistical significance (all $P = 0.114$). RETFound-DINOv2 was not re-evaluated on External Test 2 because images from AlzEye 2018 contributed to the Moorfields fundus data used in its pre-training, and we wished to avoid potential optimism arising from data reuse; therefore, only the ensemble and JOIR models are compared in this cohort.

Table 6 shows the associations between RAG and the interview-based medical conditions after age- and sex-matched propensity-score analysis in the systemic disease assessment set. The age gap for participants who were taking medication for hypertension was −0.39 ± 3.96 years, whereas it was −0.75 ± 4.34 years for those who were not taking medication, and this difference was not significant in the overall cohort ($P = 0.38$). However, in the Low-SD subgroup (≤ 1.35 years), the gap shifted significantly towards older values in medicated participants (0.04 ± 3.53 years vs. −1.00 ± 3.06 years; $P = 5.74 \times 10^{-15}$). Participants taking medication for diabetes had an age gap of 0.61 ± 4.11 years, compared to −0.70 ± 4.19 years for those who were not taking medication ($P = 1.98 \times 10^{-6}$). Individuals with a history of stroke had an age gap of 0.33 ± 3.70 years, versus −0.42 ± 2.69 years in participants without stroke ($P = 0.019$). Likewise, participants with a history of cardiac disease exhibited an age gap of −0.33 ± 3.38 years, compared to −0.78 ± 2.91 years for those without a history of cardiac disease ($P = 0.003$). In contrast, no significant differences were observed among participants receiving lipid-lowering medicines ($P = 0.367$) or those with chronic renal disease ($P = 0.101$).

Discussion

This study demonstrates that combining a simple two-output MTL architecture with ensemble learning can provide advantages over STL frameworks for estimating biological age from retinal fundus images. By sharing representations between age and HbA1c, the MTL model regularised image features and reduced age-prediction error relative to a fundus-only STL baseline, whereas an ensemble of fold-specific models further stabilised performance. This behaviour is consistent with previous work showing that MTL can exploit shared structure between related targets to mitigate overfitting and that deep ensembles reduce variance and enhance robustness[36,37]. Across internal and external cohorts, the MTL ensemble achieved accuracy comparable to, or better than, existing retinal age models while retaining an image-only inference pipeline. Consistent with clinical expectations, Grad-CAM visualisations indicated that both the fold-specific models and the ensemble primarily attended to the optic disc, macula, and major temporal vascular arcades while down-weighting peripheral artefacts such as image borders (Fig. 3), supporting the biological plausibility of the learned features.

In the broader landscape of retinal AI, foundation models and automated quality-control pipelines provide complementary routes to robustness. Self-supervised retinal foundation models such as RETFound are trained on very large collections of unlabelled retinal images and can then be efficiently adapted to diverse diagnostic and prognostic tasks with comparatively modest labelled datasets[26]. More recently, the RETFound-DINOv2 family has extended this framework by varying the pre-training data composition across large UK and Chinese retinal cohorts and systematically characterising the impact of pre-training data on generalisability and fairness in downstream oculomics applications[27]. In parallel, general-purpose vision transformers such as DINOv2, pre-trained self-supervised on 142 million natural images, have shown that high-capacity foundation models can provide strong, reusable features across visual domains[31]. In this study, we used a RETFound-DINOv2–based retinal age-prediction model as our retinal foundation-model baseline. In our primary external cohort,

both the baseline and fine-tuned RETFound-DINOv2 age-prediction models outperformed the JOIR retinal age benchmark, underscoring the strength of specialist retinal foundation models trained on large multi-site datasets[28]. By contrast, within this relatively narrow setting, our MTL ensemble achieved slightly lower error than both RETFound-DINOv2 baselines (Table 3). We suspect that this reflects the relatively narrow domain of our setting—both the development data and External Test 1 comprised Japanese health-check cohorts with similar acquisition protocols—and the fact that simple chronological age regression can already be captured well by a lightweight convolutional architecture trained from scratch. We therefore view our model as complementary to high-capacity retinal foundation models such as RETFound-DINOv2, which are designed for broader, more heterogeneous tasks and populations. Future work should explore combining such foundation models with extensive multi-camera training, data augmentation, and pre-diagnostic modules that assess image quality, field of view, and laterality to further mitigate device- and acquisition-related variability; such pipelines have been externally validated across centres and ethnicities[38-40]. More recently, multimodal visual–language and eyecare foundation models, including EyeFM, have been proposed that jointly encode retinal images and clinical or textual information to support clinical decision making[41,42]. Within this context, our two-output MTL ensemble should be viewed as a pragmatic, lightweight strategy that enhances accuracy and stability in retinal age estimation without requiring the computational or annotation resources of large foundation models, and could, in future, be embedded as an image encoder within more comprehensive multimodal systems.

A key design choice in this study was to restrict the MTL architecture to two concurrent regression targets, age and a single systemic biomarker, so that the effect of adding one auxiliary task could be evaluated clearly and the risk of negative transfer from heterogeneous outputs could be limited. In model-selection experiments, predicting HbA1c alongside age modestly reduced age-prediction error, whereas using systolic blood pressure or LDL cholesterol as the auxiliary output either conferred little

benefit or worsened performance relative to the single-task baseline. HbA1c reflects average blood glucose concentrations over approximately 2–3 months and is less susceptible to short-term fluctuations than parameters such as systolic blood pressure or LDL cholesterol[43,44]. These properties make HbA1c a stable supervisory signal for shaping retinal representations in MTL. Epidemiological studies have further shown that HbA1c levels increase with age even in non-diabetic populations, including Japanese working cohorts, supporting its role as a marker of age-related metabolic change[45,46]. Beyond HbA1c, multimodal retinal AI systems have also combined fundus images with systemic variables or laboratory measures—including kidney function markers—to detect and prognosticate chronic kidney disease and diabetic kidney disease, illustrating the broader utility of jointly modelling ocular and systemic data in deep-learning frameworks[47,48]. Taken together with our comparative experiments, these findings support the use of HbA1c as an auxiliary output during training to regularise an image-only retinal age model, so that deployment requires only a single fundus photograph.

The concept of exploiting retinal information as a non-invasive window to systemic health has deep clinical roots. Classical ophthalmoscopic signs, including arteriolar narrowing, arteriovenous nicking, and retinopathy, correlate with hypertension, atherosclerosis, and diabetes mellitus[49-52], and the Keith–Wagener–Barker classification remains a standard for grading hypertensive retinopathy[53,54]. Quantitative vascular metrics such as retinal fractal dimension have been proposed to capture vascular complexity and strengthen these associations[55]. More recently, AI-based image analysis has extended retinal–systemic links to cardiovascular risk, cognitive impairment, and neurodegenerative disorders including Alzheimer's and Parkinson's disease[11–16,56,57]. Large cohort studies have also reported that retinal ageing markers, including the RAG and related indices, predict cardiovascular events, chronic obstructive pulmonary disease, kidney failure, and Parkinson's disease, supporting their interpretation as general indicators of systemic ageing and vulnerability[13,14,17–19]. These developments parallel the broader oculomics agenda, which integrates ocular and systemic information to improve risk

stratification and disease prediction[3].

Our results are broadly consistent with this literature. In our health-check cohort, RAG values were significantly higher in participants with diabetes, prior stroke, or a history of cardiac disease, indicating a shift towards accelerated retinal ageing, whereas no association was observed for lipid-lowering medication use or chronic kidney disease. Hypertension did not reach significance in the overall cohort; however, it was significant in the subset with low ensemble SD, suggesting that stratifying eyes by ensemble disagreement can serve as a proxy for prediction confidence and may help identify subgroups in which retinal age is most informative. The direction of these associations accords with previous reports linking larger RAGs to increased risks of stroke, arterial stiffness, and incident cardiovascular disease[11–16]. At the same time, the attenuation of some associations after stratification highlights room for further refinement of retinal age models to capture more nuanced links between retinal biomarkers and systemic health.

Despite these strengths, certain limitations should be acknowledged. Our analyses relied on high-quality retinal images acquired in health-check settings, and although poor-quality photographs were excluded by expert review supported by automated checks, residual variation in focus, illumination, and fixation may still have affected performance. Most data were collected from a single geographical region, and all cohorts, including the Asian subset of the UK AlzEye screening study used for External Test 2[25], were derived from Asian populations; ethnic heterogeneity within Asian populations, together with differences in healthcare context and imaging devices, likely contributed to the larger absolute errors observed in External Test 2 and may limit generalisability to other ethnic and socioeconomic groups. The systemic disease analyses were based on self-reported diagnoses and medication use, which may contain inaccuracies and did not distinguish, for example, benign arrhythmias from prior myocardial infarction. Propensity-score matching was performed only on age and sex; thus, residual confounding by comorbidities, overall medication burden and lifestyle factors

may have biased the observed associations between RAG and systemic conditions. Biological ageing is influenced by genetics, lifestyle, and chronic disease burden, which are not fully captured by retinal imaging alone; hence, retinal age should be interpreted as one component of a broader risk profile. Furthermore, although we evaluated the ensemble in two independent external cohorts totalling more than 5,000 individuals, both cohorts comprised only Asian participants, and we did not evaluate performance in non-Asian or broadly multiethnic biobank populations; in contrast, previous retinal age studies have leveraged biobank-scale datasets to demonstrate prospective associations with incident cardiovascular events and mortality[13,14,17–19]. Model selection and hyperparameter tuning were based on cross-validation within a single institutional cohort; thus, the ensemble may still be partially overfitted to this source population despite early stopping and regularisation. Larger, fully independent and ethnically diverse external cohorts with longitudinal follow-up are warranted to more definitively establish generalisability and prognostic value. In future work, we plan to leverage the 3-year longitudinal cohort described in this study to examine how RAG trajectories relate to incident systemic disease.

From a clinical perspective, the most immediate use case for our model is as a low-friction, image-only screening tool in settings where fundus photography is already performed for ocular evaluation. Individuals with markedly positive RAG values and low ensemble SD could be flagged during routine health checks for more detailed cardiometabolic assessment, whereas near-zero RAG in the context of normal systemic risk factors might provide additional reassurance regarding vascular ageing. Prospective studies are warranted to define clinically meaningful thresholds, determine whether acting on retinal age gap–based alerts improves outcomes, and clarify how retinal age models can complement multimodal retinal-phenotype frameworks such as RetiAGE and related indices[13,14,17–19]. In conclusion, accurate estimation of retinal age in healthy populations using AI supports its use as a biomarker for systemic disease and aligns with the broader aim of integrating retinal biomarkers with

other clinical indicators for personalised risk stratification and disease prevention[13,14,17–19].

References

1. Ting, D.S.W. et al. Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol.* **103**, 167–175 (2019).

2. Aggarwal, R. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit. Med.* **4**, 65 (2021).

3. Wagner, S.K. et al. Insights into systemic disease through retinal imaging-based oculomics. *Transl. Vis. Sci. Technol.* **9**, 6 (2020).

4. Zhu, Z. et al. Oculomics: current concepts and evidence. *Prog. Retin. Eye Res.* **106**, 101350 (2025).

5. Patterson, E.J. et al. Oculomics: a crusade against the four horsemen of chronic disease. *Ophthalmol. Ther.* **13**, 1427–1451 (2024).

6. An, G., Akiba, M., Omodaka, K., Nakazawa, T. & Yokota, H. Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images. *Sci. Rep.* **11**, 4250 (2021).

7. Yim, J. et al. Predicting conversion to wet age-related macular degeneration using deep learning.

Nat. Med. **26**, 892–899 (2020).

8. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).

9. Ting, D.S.W. 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* **318**, 2211–2223 (2017).

10. Li, Z. et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on colour fundus photographs. *Ophthalmology* **125**, 1199–1206 (2018).

11. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).

12. Chen, R. et al. Association between cardiovascular health metrics and retinal ageing. *Geroscience* **45**, 1511–1521 (2023).

13. Nusinovici, S. et al. Retinal photograph-based deep learning predicts biological age, and stratifies morbidity and mortality risk. *Age Ageing* **51**, afac065 (2022).

14. Zhu, Z. et al. Retinal age gap as a predictive biomarker for mortality risk. *Br. J. Ophthalmol.* **107**, 547–554 (2023).

15. Zhu, Z. et al. Retinal age gap as a predictive biomarker of stroke risk. *BMC Med.* **20**, 466 (2022).

16. Zhu, Z. et al. Association of retinal age gap with arterial stiffness and incident cardiovascular disease. *Stroke* **53**, 3320–3328 (2022).

17. Zhang, S. et al. Association of retinal age gap and risk of kidney failure: a UK Biobank study. *Am. J. Kidney Dis.* **82**, 410–419.e1 (2023).

18. Hu, W. et al. Retinal age gap as a predictive biomarker of future risk of Parkinson's disease. *Age Ageing* **51**, afac062 (2022).

19. Peng, Q. et al. Predictive potential of retina-based biological age in assessing chronic obstructive pulmonary disease risk. *Clin. Exp. Ophthalmol.* **53**, 402–408 (2025).

20. Grimbly, M.J. et al. Estimating biological age from retinal imaging: a scoping review. *BMJ Open Ophthalmol.* **9**, e001794 (2024).

21. Liu, R. et al. Ocular ageing biomarkers and their clinical utility: a review. *Vis. Neurosci.* **42**,

e009 (2025).

22. Yu, Z. et al. A cross population study of retinal aging biomarkers with longitudinal pre-training and label distribution learning. *NPJ Digit. Med.* **8**, 344 (2025).

23. Ahadi, S. et al. Longitudinal fundus imaging and its genome-wide association analysis provide evidence for a human retinal aging clock. *Elife* **12**, e82364 (2023).

24. Gonçalves, M.B. et al. Image quality assessment of retinal fundus photographs for diabetic retinopathy in the machine learning era: a review. *Eye* **38**, 426–433 (2024).

25. Wagner, S.K. et al. AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353,157 patients in London, UK. *BMJ Open* **12**, e058552 (2022).

26. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).

27. Zhou, Y. et al. Revealing the impact of pre-training data on medical foundation models. Preprint at Research Square (2025). DOI: 10.21203/rs.3.rs-6080254/v1.

28. Miyake, M., Akiyama, M., Kashiwagi, K., Sakamoto, T. & Oshika, T. Japan Ocular Imaging

Registry: a national ophthalmology real-world database. *Jpn. J. Ophthalmol.* **66**, 499–503 (2022).

29. Sharma, P. et al. A lightweight deep learning model for automatic segmentation and analysis of ophthalmic images. *Sci. Rep.* **12**, 8505 (2022).

30. Selvaraju, R.R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359 (2020).

31. Oquab, M. et al. DINOv2: Learning robust visual features without supervision. Preprint at arXiv:2304.07193 (2023).

32. Chen, X. et al. Symbolic Discovery of Optimization Algorithms. Preprint at https://ui.adsabs.harvard.edu/abs/2023arXiv230206675C (2023).

33. Zou, H. & Hastie, T. Regularization and variable selection via the Elastic Net. *J. R. Stat. Soc. Series B Stat. Methodol.* **67**, 301–320 (2005).

34. Saleh, R.A. & Ehsanes Saleh, A.K.M. Statistical properties of the log-cosh loss function used in machine learning. Preprint at https://ui.adsabs.harvard.edu/abs/2022arXiv220804564S (2022).

35. Ninomiya, T. LWBNA-MTL (GitHub repository). Zenodo. https://doi.org/10.5281/zenodo.18815076 (2026).

36. Caruana, R. Multitask learning. *Mach. Learn*. **28**, 41–75 (1997).

37. Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M. & Suganthan, P.N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **115**, 105151 (2022).

38. Asaoka, R. et al. Validation of a deep learning model to screen for glaucoma using images from different fundus cameras and data augmentation. *Ophthalmol. Glaucoma* **2**, 224–231 (2019).

39. Asaoka, R. et al. Usefulness of data augmentation for visual field trend analyses in patients with glaucoma. *Br. J. Ophthalmol.* **104**, 1697–1703 (2020).

40. Yuen, V. et al. Deep-learning-based pre-diagnosis assessment module for retinal photographs: a multicenter study. *Transl. Vis. Sci. Technol.* **10**, 16 (2021).

41. Shi, D. et al. A multimodal visual–language foundation model for computational ophthalmology. *NPJ Digit. Med.* **8**, 381 (2025).

42. Wu, Y. et al. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nat. Med.* **31**, 3404–3413 (2025).

43. Garg, P., Pethusamy, K. & Ranjan, R. Correlation between estimated average glucose levels

calculated from HbA1c values and random blood glucose levels in a cohort of subjects. *J. Lab. Physicians* **15**, 217–223 (2023).

44. Fatica, E.M. et al. Short- and long-term biological variability of small dense LDL, HDL3, and triglyceride-rich lipoprotein cholesterol. *J. Appl. Lab. Med.* **7**, 1047–1061 (2022).

45. Pani, L.N. et al. Effect of aging on A1C levels in individuals without diabetes: evidence from the Framingham Offspring Study and the National Health and Nutrition Examination Survey 2001–2004. *Diabetes Care* **31**, 1991–1996 (2008).

46. Hashimoto, Y., Futamura, A., & Ikushima, M. Effect of aging on HbA1c in a working male Japanese population. *Diabetes Care* **18**, 1337–1340 (1995).

47. Zhang, K. et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nat. Biomed. Eng.* **5**, 533–545 (2021).

48. Meng, Z. et al. Non-invasive biopsy diagnosis of diabetic kidney disease via deep learning applied to retinal images: a population-based study. *Lancet Digit. Health* **7**, 100868 (2025).

49. Kim, D.H. Retinal microvascular signs and disability in the Cardiovascular Health Study. *Arch. Ophthalmol.* **130**, 350 (2012).

50. Wang, X.Y. et al. Retinal vascular morphological changes in patients with extremely severe obstructive sleep apnea syndrome. *Chin. Med. J. (Engl.)* **130**, 805–810 (2017).

51. Benitez-Aguirre, P. et al. Retinal vascular geometry predicts incident retinopathy in young people with type 1 diabetes. *Diabetes Care* **34**, 1622–1627 (2011).

52. Cheung, C.Y.-L., Ikram, M.K., Sabanayagam, C. & Wong, T.Y. Retinal microvasculature as a model to study the manifestations of hypertension. *Hypertension* **60**, 1094–1103 (2012).

53. Li, L.J., Ikram, M.K. & Wong, T.Y. Retinal vascular imaging in early life: insights into processes and risk of cardiovascular disease. *J. Physiol.* **594**, 2175–2203 (2016).

54. Keith, N.M., Wagener, H.P. & Barker, N.W. Some different types of essential hypertension: their course and prognosis. *Am. J. Med. Sci.* **268**, 336–345 (1974).

55. Engelmann, J. et al. Retinal fractal dimension is a potential biomarker for systemic health—evidence from a mixed-age, primary-care population. *Transl. Vis. Sci. Technol.* **13**, 19 (2024).

56. Wagner, S.K. et al. Retinal optical coherence tomography features associated with incident and

prevalent Parkinson disease. *Neurology* **101**, e1581–e1593 (2023).

57. Shao, Y. et al. Visualization of focal thinning of the ganglion cell-inner plexiform layer in patients with mild cognitive impairment and Alzheimer's disease. *J. Alzheimers Dis.* **64**, 1261–1273 (2018).

Acknowledgments

This research was partly supported by the Japan Science and Technology Agency (JST) through the Center of Innovation Program (COI-NEXT), grant no. JPMJPF2201. The funder had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Author Contributions

T.N.I. conceived and designed the study, implemented the code, analysed the data and wrote the manuscript. A.H. contributed to study design and critically revised the manuscript. N.K., P.S., Y.Z., S.W., K.S., T.NO., N.T., K.O., and N.H. advised on methodology and data analysis. T.M. collected the fundus images and constructed the datasets. Y.I., P.K., and T.N. supervised the project and refined the overall experimental design. All the authors discussed the results and contributed to reviewing and editing the final manuscript.

Competing Interests

The authors declare no conflict of interest.

Data Availability

The retinal images and associated clinical records analysed in this study contain personally identifiable

medical information and are therefore not publicly available. The datasets used in this study include the development and internal-validation datasets, an external-validation dataset from Sendai Open Hospital, and the AlzEye dataset. De-identified data from the development and internal-validation datasets may be provided by the corresponding author (Toru Nakazawa) upon reasonable request, subject to approval by the Tohoku University Institutional Review Board and the signing of a data-sharing agreement; we aim to respond to requests within 30 days. The external-validation dataset from Sendai Open Hospital is not publicly available and may be provided upon reasonable request, subject to appropriate institutional approvals and a data-sharing agreement. The AlzEye dataset is available via the dataset providers under their data access and governance procedures. Source data underlying the graphs and charts presented in the main figures are provided as Supplementary Data 1. Study data are stored on secure, access-controlled servers at Tohoku University Graduate School of Medicine.

Code Availability

Custom code and reproducible scripts used for model training and data analysis are publicly available at https://github.com/stereocord/LWBNA-MTL/ and are archived on Zenodo (DOI: 10.5281/zenodo.18815076)[35]. Software versions and key parameters used in this study are described in the Methods section.

Figure legends

Fig. 1. Inclusion and exclusion criteria for creating the training, internal-validation, and systemic disease assessment datasets from retinal imaging records. Flowchart depicting the inclusion and exclusion criteria utilised to create the study dataset. Of the 64,772 individuals (157,695 eye records)

who underwent health screening, those with missing clinical data, poor image quality, and specific medical conditions or medications were excluded. This process yielded 27,214 individuals with 50,595 eye records for the training set and 7,288 individuals with 7,288 eye records for the internal-validation set. In parallel, 8,467 individuals (8,467 eyes) were retained as the systemic disease assessment set.

Fig. 2. Schematic architectures of the single-task learning (STL) and multitask learning (MTL) models. a) STL model. A convolutional neural network (CNN) processes each fundus image; the resulting feature vector passes through a single fully connected (fc) head that regresses chronological age. **b**) MTL model. The same CNN backbone feeds two parallel fc heads that simultaneously predict age and a second biometric parameter (e.g., HbA1c). Joint optimisation of the two outputs updates the shared weights differently from the STL model. Fully connected layers (fc) are the dense components of a neural network that integrate and transform the CNN-extracted features into the final regression outputs. The CNN backbone in both configurations is a modified LWBNA_Unet with residual blocks, as detailed in Methods.

Fig. 3. Visualisation of age-prediction outputs and saliency across multiple folds and the multitask learning (MTL) ensemble model. The output of the individual fold models and the ensemble model for age prediction is shown, wherein the top row presents Grad-CAM heat maps that highlight the key retinal regions used for the prediction by each model, whereas the bottom row displays the predicted ages overlaid on the original images. From left to right, the images represent the results from Fold 1 to Fold 5, followed by the ensemble model. The ensemble combines the five fold-specific models, yielding more accurate and consistent age estimates than any single-fold models. By pooling their diverse attention maps, it sharpens focus on the shared anatomical landmarks (optic disc, macula, and major vessels), thereby increasing robustness.

Fig. 4. Relationship between actual and predicted age by the multitask learning (MTL) ensemble model. a) Bland–Altman plot showing the agreement between actual and predicted ages by the MTL ensemble model in the internal validation cohort (n = 7,288 eyes). The majority of data points fell within the ±1.96 standard deviation (SD) limits, indicating good agreement between predicted and actual ages. **b**) Scatterplot illustrating the relationship between actual age and predicted age. The strong correlation is demonstrated by the Pearson correlation coefficient ($r = 0.97$, $P < 1 \times 10^{-300}$), and the regression line closely follows the line of equality, which indicates a high degree of accuracy in the model's predictions.

Fig. 5. Stratification of prediction error by the standard deviation (SD) of ensemble outputs. a) Scatterplot of absolute age-prediction error versus predicted age SD in the internal-validation cohort (n = 7,288 eyes). The vertical green line marks the cohort-median SD (1.71 years), separating the Low-SD subgroup (blue) from the High-SD subgroup (red).

b) Box-and-whisker plots comparing absolute errors between the two subgroups; boxes show the median and interquartile range (IQR), and whiskers extend to 1.5 × IQR. A two-sided Wilcoxon rank-sum test indicates a significant difference ($P = 5.95 \times 10^{-22}$).

c) Performance summary for each subgroup (mean ± SD). The Low-SD group exhibits lower MAE, MSE and MAPE than the High-SD group.

Tables

Table 1. Baseline demographic and clinical characteristics of the training, internal-validation, and external-validation cohorts

Variable	(A) Train	(B) Internal Test	(C) External Test 1	(D) External Test 2	A vs. B *(p ; |SMD|)*	A vs. C *(p ; |SMD|)*	A vs. D *(p ; |SMD|)*
Individuals, n	27,214	7,288	135	4,992	-	-	-
Eyes, n	50,595	7,288	135	4,992	-	-	-
Age, years	52.5 ± 14.7 (17–94)	50.5 ± 15.3 (16–91)	53.1 ± 9.6 (25–69)	57.0 ± 12.8 (30–95)	**2.34×10^{-23}** ; 0.134	0.442 ; 0.044	**1.16×10^{-106}** ; 0.312
Male sex, n (%)	13,382 (49.2)	4,030 (55.3)	113 (83.7)	2,582 (51.7)	**1.82×10^{-25}** ; 0.123	**2.09×10^{-18}** ; 0.786	**9.27×10^{-4}** ; 0.051
Systolic BP, mmHg	127.9 ± 21.3	124.8 ± 20.5	121.5 ± 14.2	-	**4.17×10^{-16}** ; 0.147	**1.11×10^{-8}** ; 0.301	-
Diastolic BP, mmHg	78.3 ± 13.0	76.7 ± 12.9	70.1 ± 11.4	-	**1.71×10^{-18}** ; 0.123	**3.94×10^{-27}** ; 0.631	-
HDL cholesterol, mmol/L	1.66 ± 0.43	1.68 ± 0.43	1.49 ± 0.40	-	**3.64×10^{-5}** ; 0.046	**1.18×10^{-11}** ; 0.410	-
LDL cholesterol, mmol/L	3.29 ± 0.86	3.23 ± 0.83	3.31 ± 0.88	-	**2.41×10^{-6}** ; 0.082	0.919 ; 0.009	-
HbA1c, %	5.7 ± 0.6	5.6 ± 0.5	5.7 ± 0.4	-	**5.13×10^{-15}** ; 0.162	0.918 ; 0.009	-
Smoking, n (%)	5,409 (19.9)	1,395 (19.1)	35 (25.9)	-	0.161 ; 0.019	0.079 ; 0.144	-

Data are shown as mean ± SD for continuous variables and as n (%) for categorical variables.

P-values for A vs. B and A vs. C were obtained using the Mann–Whitney *U* or χ^2 tests.

For A vs. D, direct raw-data comparison was not allowed because external dataset could not be transferred across institutions; *P*-values were therefore estimated from summary statistics using Welch's two-sample t-test for continuous variables and the χ^2 test for categorical variables. Absolute standardised mean differences (|SMD|) versus the training cohort (A) are shown in the last three columns; all pairwise P-values and |SMD| values, including B–C, B–D and C–D comparisons, are provided in Supplementary Table 1. Boldface indicates statistically significant *P*-values ($P < 0.05$).

Abbreviations: BP, blood pressure; HbA1c, glycated haemoglobin; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SD, standard deviation.

Table 2. Age-prediction performance of the single-task and multitask models under different input–output configurations

Type	Input variables	Output variables	5-fold MAE (years)	5-fold MSE (years^2)	5-fold MAPE (%)	Ensemble MAE (years)	ΔMAE (vs. baseline, years)	*p* (vs. baseline)
STL (baseline)	Fundus	Age	5.47 ± 0.50	53.66 ± 9.05	12.52 ± 1.31	5.03	-	-
STL	Fundus + HbA1c	Age	5.86 ± 0.47	61.06 ± 10.10	13.21 ± 1.09	5.19	+0.15	**3.21×10^{-5}**
STL	Fundus + Sex	Age	5.52 ± 0.10	53.89 ± 2.36	12.61 ± 0.29	5.15	+0.11	**4.12×10^{-4}**
MTL	Fundus + Sex	Age + HbA1c	5.44 ± 0.17	52.69 ± 2.25	12.40 ± 0.61	5.05	+0.02	0.284
MTL	Fundus	Age + HbA1c	5.31 ± 0.33	49.62 ± 5.81	12.12 ± 0.90	4.86	-0.18	**1.27×10^{-4}**
MTL	Fundus	Age + SBP	5.91 ± 0.60	60.64 ± 12.25	13.73 ± 1.75	5.40	+0.37	**5.68×10^{-6}**
MTL	Fundus	Age + LDL	5.90 ± 0.32	61.81 ± 6.42	13.83 ± 1.11	5.46	+0.43	**2.41×10^{-6}**

Values are mean ± SD across the five fold-specific models trained by five-fold cross-validation on a reduced training subset (approximately 5% random subsample of eye records from the training cohort) and evaluated on the same independent internal-validation cohort (n = 7,288). “Ensemble MAE” denotes the mean absolute error obtained by aggregating the five fold-specific models via arithmetic averaging of their predictions on the same internal-validation eyes. For each candidate model, ΔMAE and *P*-values are calculated relative to the baseline STL model (fundus-only age prediction) using two-sided Wilcoxon signed-rank tests on per-eye absolute errors derived from the ensemble predictions. For multitask-learning (MTL) models, performance metrics refer to the age-prediction output; the auxiliary systemic variable (for example, HbA1c) was not used as a primary selection target. Boldface indicates statistically significant *P*-values ($P < 0.05$).

Abbreviations: HbA1c, glycated haemoglobin; LDL, low-density lipoprotein; MAE, mean absolute error; MAPE, mean absolute percentage error; MSE, mean squared error; MTL, multitask learning;
SBP, systolic blood pressure; SD, standard deviation; STL, single-task learning.

Table 3. External-validation comparison between the MTL ensemble model and a benchmark retinal age model

Subgroup	Metric	MTL ensemble (A)	JOIR (B)	RETFound (fine-tuned, C)	RETFound (baseline, D)	*P-value* A vs. B	*P*-value A vs. C	*P*-value A vs. D
Overall	MAE (years)	3.39 ± 2.74	5.37 ± 3.84	4.45 ± 3.15	4.65 ± 3.31	**3.17×10^{-7}**	**0.002**	**1.04×10^{-4}**
(n = 135 eyes)	MSE ($years^2$)	18.95 ± 30.48	43.57 ± 58.50	29.77 ± 37.35	32.58 ± 40.60	**1.16×10^{-7}**	**0.002**	**4.01×10^{-5}**
	MAPE (%)	6.60 ± 5.59	10.15 ± 7.00	8.48 ± 5.83	8.78 ± 6.37	**9.70×10^{-7}**	**0.003**	**2.01×10^{-4}**
Low-SD	MAE	2.87 ± 2.14	5.07 ± 3.12	4.06 ± 3.09	4.02 ± 2.83	**2.70×10^{-5}**	**0.018**	**0.011**
(≤ 2.35 years, n = 68)	MSE	12.81 ± 18.16	35.50 ± 38.22	25.98 ± 35.40	24.23 ± 29.64	**2.77×10^{-5}**	**0.019**	**0.008**
	MAPE	5.58 ± 4.48	9.75 ± 5.96	7.91 ± 5.97	7.65 ± 5.11	**5.51×10^{-5}**	**0.021**	**0.012**
High-SD	MAE	3.91 ± 3.15	5.67 ± 4.43	4.86 ± 3.16	5.29 ± 3.62	**0.002**	**0.044**	**0.003**
(> 2.35 years, n = 67)	MSE	25.18 ± 38.22	51.76 ± 72.66	33.62 ± 38.85	41.07 ± 47.83	**6.33×10^{-4}**	0.052	**0.002**
	MAPE	7.64 ± 6.36	10.56 ± 7.89	9.07 ± 5.62	9.93 ± 7.25	**0.003**	0.053	**0.007**

Data are mean ± SD. Low-SD and high-SD subgroups are defined by the median ensemble SD (2.35 years). *P*-values are from two-sided Wilcoxon signed-rank tests comparing paired predictions of the multitask learning (MTL) ensemble and the benchmark model on the same eyes (n = 135). Boldface indicates statistically significant *P*-values ($P < 0.05$).

Abbreviations: MAE, mean absolute error; MAPE, mean absolute percentage error; MSE, mean squared error; MTL, multitask learning; SD, standard deviation.

Table 4. External-validation comparison between the MTL ensemble model and a benchmark retinal age model

Subgroup	Metric	MTL ensemble	JOIR model	*P-value* (A vs. B)
Overall (n = 4,992 eyes)	MAE (years)	8.63 ± 6.33	9.02 ± 7.55	**0.003**
	MSE (years^2)	114.56 ± 166.88	138.40 ± 236.27	**0.003**
	MAPE (%)	16.52 ± 13.78	16.96 ± 15.69	**0.003**
Low-SD (≤ 2.68 years, n = 2,496)	MAE	8.10 ± 5.94	8.41 ± 6.88	**0.008**
	MSE	100.90 ± 144.66	118.01 ± 191.82	**0.008**
	MAPE	14.95 ± 13.11	15.20 ± 14.25	**0.008**
High-SD (> 2.68 years, n = 2,496)	MAE	9.17 ± 6.65	9.63 ± 8.12	0.114
	MSE	128.23 ± 185.50	158.80 ± 272.10	0.114
	MAPE	18.08 ± 14.25	18.72 ± 16.83	0.114

Data are mean ± SD. Low-SD and high-SD subgroups are defined by the median ensemble standard deviation (2.68 years). *P*-values are from two-sided Wilcoxon signed-rank tests comparing paired predictions of the multitask learning (MTL) ensemble and the JOIR model on the same eyes (n = 4,992). Boldface indicates statistically significant *P*-values ($P < 0.05$).

Abbreviations: MAE, mean absolute error; MAPE, mean absolute percentage error; MSE, mean squared error; MTL, multitask learning; SD, standard deviation.

Table 5. Internal cross-validation performance of fold-specific MTL models and their ensemble

Model (five-fold CV)	MAE (years)	MSE (years^2)	MAPE (%)
Fold 1 (single MTL)	3.11 ± 2.71	17.02 ± 36.35	6.60 ± 6.01
Fold 2 (single MTL)	3.12 ± 2.61	16.52 ± 30.74	6.68 ± 5.90
Fold 3 (single MTL)	2.99 ± 2.53	15.34 ± 29.74	6.44 ± 6.02
Fold 4 (single MTL)	3.01 ± 2.60	15.80 ± 31.71	6.38 ± 5.90
Fold 5 (single MTL)	2.97 ± 2.56	15.36 ± 32.68	6.18 ± 5.41
Ensemble (mean of folds)	2.78 ± 2.39	13.42 ± 27.47	5.88 ± 5.36

Each “Fold” row reports the performance of a single multitask learning (MTL) model trained on four folds of the training cohort and evaluated on the same internal-validation cohort (n = 7,288). Reported values are mean ± SD calculated directly from the distribution of prediction errors for all eyes in that internal-validation cohort. “Ensemble” aggregates the five fold-specific models by arithmetic averaging of their predictions on that internal-validation set, enabling direct comparison with the single-model results.

Abbreviations: CV, cross-validation; MAE, mean absolute error; MAPE, mean absolute percentage error; MSE, mean squared error; MTL, multitask learning; SD, standard deviation.

Table 6. Association between retinal age gap and self-reported systemic conditions

Condition	n All/Low/High	Metric	Overall retinal age gap (years)			Low-SD subgroup (≤ 1.35 y)			High-SD subgroup (> 1.35 y)		
			Yes	No	*P*-value	Yes	No	*P*-value	Yes	No	*P*-value
Medication for hypertension	3,718/1,746/1,972	RAG	-0.39 ± 3.96	-0.75 ± 4.34	0.38	0.04 ± 3.53	-1.00 ± 3.06	**5.74×10^{-15}** §	-0.76 ± 4.27	-0.93 ± 3.41	**0.032**†
		MAE	3.00 ± 2.61	3.23 ± 2.98	0.485	2.64 ± 2.35	2.45 ± 2.09	0.215	3.33 ± 2.79	2.83 ± 2.11	**0.003**‡
Medication for diabetes	840/404/436	RAG	0.61 ± 4.11	-0.70 ± 4.19	**1.98×10^{-6}** §	0.85 ± 3.53	-0.22 ± 2.87	**7.38×10^{-4}** §	0.39 ± 4.58	-1.38 ± 3.69	**4.27×10^{-7}** §
		MAE	3.14 ± 2.72	3.33 ± 2.63	0.233	2.81 ± 2.29	2.14 ± 1.92	**0.002**‡	3.44 ± 3.03	2.89 ± 2.68	0.072
Lipid-lowering medications	2,450/1,192/1,258	RAG	-0.57 ± 3.93	-0.50 ± 3.75	0.367	-0.03 ± 3.45	-0.11 ± 2.72	0.694	-1.09 ± 4.27	-1.44 ± 3.15	0.05
		MAE	3.02 ± 2.57	2.93 ± 2.40	0.584	2.65 ± 2.20	1.97 ± 1.87	**4.08×10^{-10}** §	3.37 ± 2.84	2.60 ± 2.28	**2.27×10^{-7}** §
History of stroke	290/134/156	RAG	0.33 ± 3.70	-0.42 ± 2.69	**0.019**†	0.22 ± 3.57	-0.46 ± 3.44	0.191	0.42 ± 3.82	0.43 ± 3.07	0.973
		MAE	2.75 ± 2.48	2.01 ± 1.83	**0.013**†	2.53 ± 2.51	2.69 ± 2.17	0.516	2.94 ± 2.46	2.47 ± 1.86	0.334
History of Cardiac Disease	980/478/502	RAG	-0.33 ± 3.38	-0.78 ± 2.91	**0.003**‡	-0.11 ± 3.10	0.00 ± 2.83	0.628	-0.55 ± 3.63	-0.83 ± 3.70	0.258
		MAE	2.57 ± 2.22	2.39 ± 1.84	0.709	2.35 ± 2.01	2.17 ± 1.81	0.568	2.78 ± 2.39	3.05 ± 2.25	0.059
History of Chronic Renal disease	118/58/60	RAG	0.84 ± 4.45	-1.02 ± 4.05	0.101	0.70 ± 4.09	-0.46 ± 3.15	0.419	0.97 ± 4.84	-0.59 ± 3.80	0.438
		MAE	3.18 ± 3.20	2.96 ± 2.92	0.667	3.28 ± 2.47	2.55 ± 1.83	0.312	3.07 ± 3.83	2.74 ± 2.65	0.947

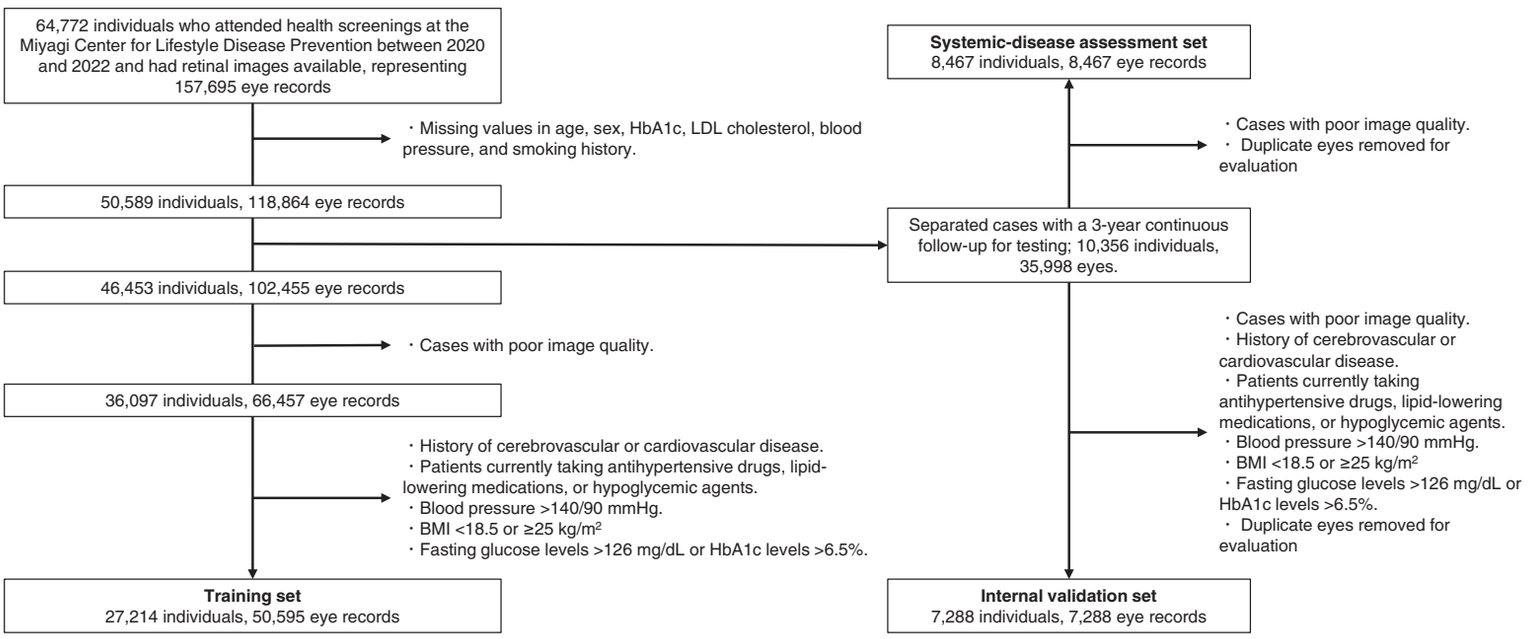
Values are mean ± SD. For RAG rows, values represent the retinal age gap (predicted age – chronological age); for MAE rows, values represent the absolute prediction error. Negative values indicate a younger-appearing retina. Low-SD and high-SD subgroups are defined by the median ensemble SD in this systemic disease cohort (1.35 years). For each condition, participants without the condition were selected by 1:1 nearest-neighbour propensity-score matching on age and sex. Boldface indicates statistically significant *P*-values ($P < 0.05$). † $P < 0.05$, ‡ $P < 0.01$, § $P < 0.001$ (two-sided Wilcoxon rank-sum tests comparing matched Yes vs. No groups).

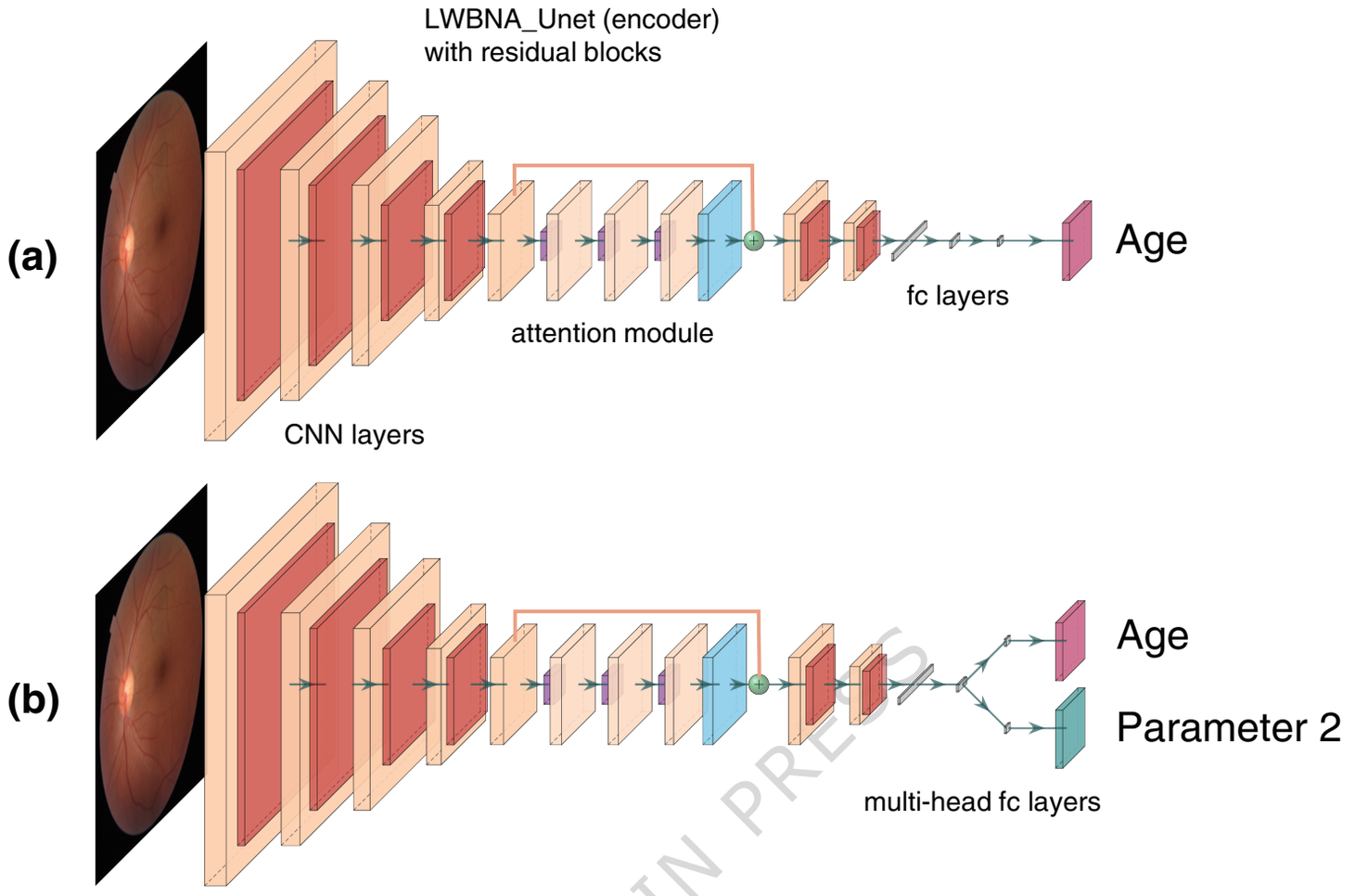
Abbreviations: MAE, mean absolute error; RAG, retinal age gap; SD, standard deviation.

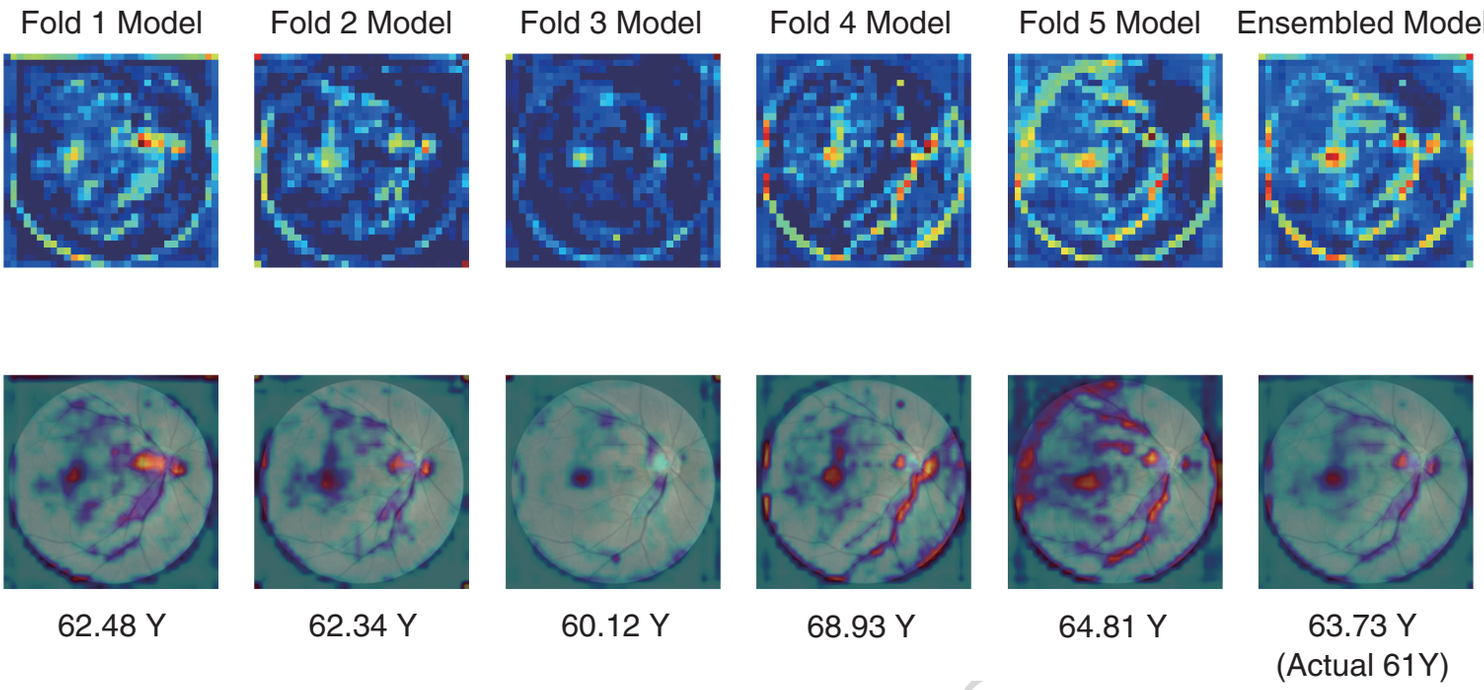
Editorial summary:

Ninomiya et al. develop an ensemble multitask deep-learning model that estimates retinal age from a fundus photograph while using an auxiliary blood marker only during training. They showed that ensemble disagreement provides confidence score and that a larger retinal age gap is associated with diabetes, cardiac disease, and stroke.

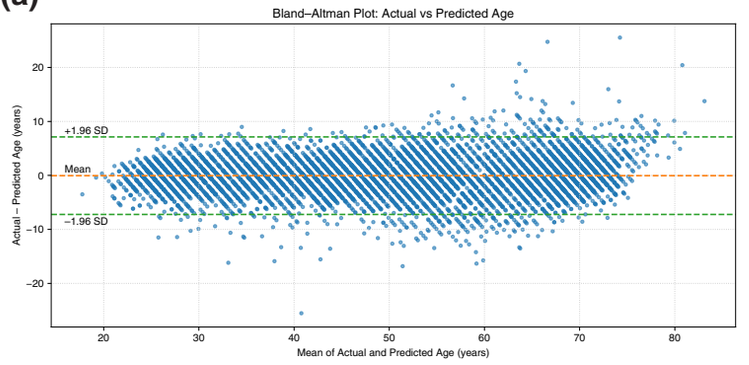
Peer review information: *Communications Medicine* thanks Jianwei Lin, Ziyao Meng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.


64,772 individuals who attended health screenings at the Miyagi Center for Lifestyle Disease Prevention between 2020 and 2022 and had retinal images available, representing 157,695 eye records
・Missing values in age, sex, HbA1c, LDL cholesterol, blood pressure, and smoking history.
50,589 individuals, 118,864 eye records
46,453 individuals, 102,455 eye records
・Cases with poor image quality.
36,097 individuals, 66,457 eye records
・History of cerebrovascular or cardiovascular disease.
・Patients currently taking antihypertensive drugs, lipid-lowering medications, or hypoglycemic agents.
・Blood pressure >140/90 mmHg.
・BMI <18.5 or ≥25 kg/m²
・Fasting glucose levels >126 mg/dL or HbA1c levels >6.5%.
Training set
27,214 individuals, 50,595 eye records
Systemic-disease assessment set
8,467 individuals, 8,467 eye records
・Cases with poor image quality.
・ Duplicate eyes removed for evaluation
Separated cases with a 3-year continuous follow-up for testing; 10,356 individuals, 35,998 eyes.
・Cases with poor image quality.
・History of cerebrovascular or cardiovascular disease.
・Patients currently taking antihypertensive drugs, lipid-lowering medications, or hypoglycemic agents.
・Blood pressure >140/90 mmHg.
・BMI <18.5 or ≥25 kg/m²
・Fasting glucose levels >126 mg/dL or HbA1c levels >6.5%.
・ Duplicate eyes removed for evaluation
Internal validation set
7,288 individuals, 7,288 eye records

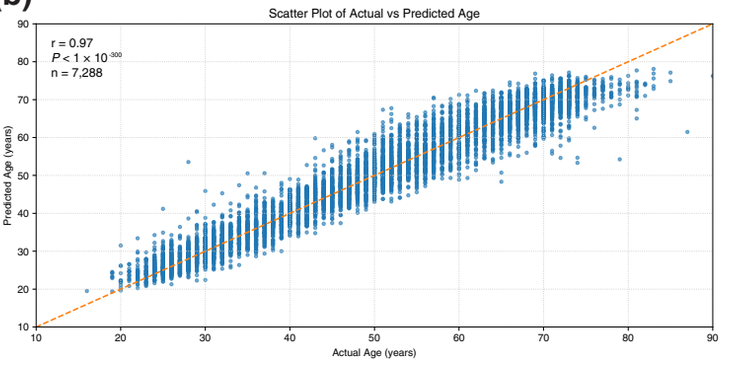




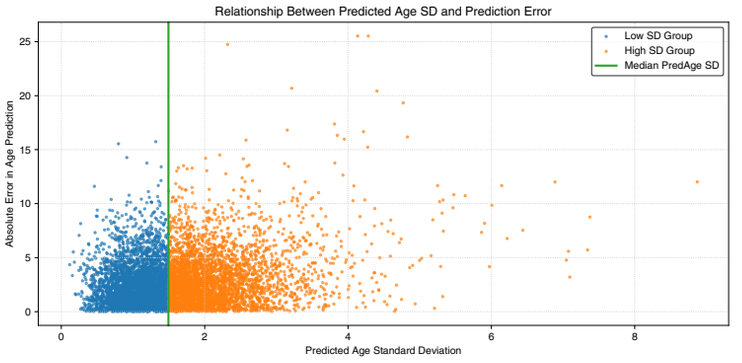
(a)



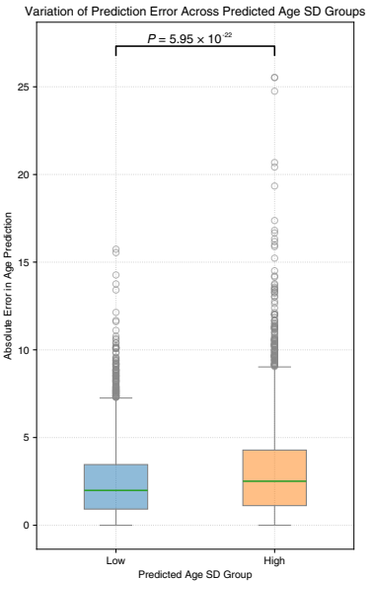
(b)



(a)



(b)



(c)

Group	MAE (years)	MSE (years^2)	MAPE (%)
Low-SD (n=3644)	2.46 ± 2.03	10.19 ± 17.48	5.37 ± 4.67
High-SD (n=3644)	3.09 ± 2.66	16.66 ± 34.39	6.39 ± 5.94