PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY OPEN ACCESS

Accurate estimation of biological age and its application in disease prediction using a multimodal image Transformer system

Jinzhuo Wang[a,1], Yuanxu Gao[b,1], Fangfei Wang[b,c], Simiao Zeng[d], Jiahui Li[d], Hanpei Miao[e], Taorui Wang[d], Jin Zeng[c], Daniel Baptista-Hon[b], Olivia Monteiro[b], Taihua Guan[c], Linling Cheng[b], Yuxing Lu[a], Zhengchao Luo[a], Ming Li[f], Jian-kang Zhu[g], Sheng Nie[h,2], Kang Zhang[a,b,c,e,2], and Yong Zhou[i,2]

Edited by Helen Mayberg, Icahn School of Medicine at Mount Sinai, New York, NY; received June 9, 2023; accepted October 12, 2023

Aging in an individual refers to the temporal change, mostly decline, in the body's ability to meet physiological demands. Biological age (BA) is a biomarker of chronological aging and can be used to stratify populations to predict certain age-related chronic diseases. BA can be predicted from biomedical features such as brain MRI, retinal, or facial images, but the inherent heterogeneity in the aging process limits the usefulness of BA predicted from individual body systems. In this paper, we developed a multimodal Transformer–based architecture with cross-attention which was able to combine facial, tongue, and retinal images to estimate BA. We trained our model using facial, tongue, and retinal images from 11,223 healthy subjects and demonstrated that using a fusion of the three image modalities achieved the most accurate BA predictions. We validated our approach on a test population of 2,840 individuals with six chronic diseases and obtained significant difference between chronological age and BA (AgeDiff) than that of healthy subjects. We showed that AgeDiff has the potential to be utilized as a standalone biomarker or conjunctively alongside other known factors for risk stratification and progression prediction of chronic diseases. Our results therefore highlight the feasibility of using multimodal images to estimate and interrogate the aging process.

biological age prediction | multimodal fusion | transformer with cross-attention | chronic disease diagnosis and prognosis | biomarker discovery

Significance

The aging process is inevitable and is a risk factor for chronic diseases. The biological age (BA) of each individual contains structural and functional determinants of aging, and its difference (AgeDiff) from the chronological age (CA) can be used as a biomarker for accelerated aging caused by underlying pathologies. We described a multimodal Transformer–based architecture which can estimate BA based on facial, fundus, and tongue images. Our results demonstrated that we can accurately estimate BA of healthy individuals, significant deviations of AgeDiff are present in individuals with chronic diseases, and AgeDiff can be used to accurately detect systematic diseases and identify progression risks. Our study highlights an approach to use easily and readily acquired patient data to identify chronic diseases.

Aging is a risk factor for many chronic diseases. However, the identification of suitable predictors of universal aging for use in health management and clinical practice has been difficult (1, 2). This is likely due to the heterogeneous nature of the underlying tissues and organ vulnerabilities associated with aging that is not simply restricted to the passage of time. Biological age (BA) on the other hand takes into account the impact of structural and functional changes that contribute to aging (3). These could be influenced by genetic and/or environmental factors. Thus, the ability to quantify BA may be clinically important to identify patients at risk for age-related diseases and raises the possibility for early intervention. AI approaches have been developed to predict BA from a number of biomarkers of aging, such as leukocyte telomere length (4), DNA methylation–based epigenetic clock (5), brain image–derived brain age (6, 7), retinal age (8, 9), and facial age (10, 11). The retina in particular has been recognized as a window to the brain due to the presence of central nervous system–derived axons in the optic nerve, as well as similarities in the expression of cytokines and immune modulators (12). The retinal age gap, the difference in the predicted retinal age and the chronological age (CA), has been used to assess brain health (13, 14). Facial age has also emerged as a potential predictor for skin health (10, 11, 15). However, while estimation of the BA of specific organs or systems may be useful to derive information regarding organ-specific diseases, utilization of BA to its full potential will undoubtedly need to take into account the heterogeneous nature of aging. We argue that the modelling of the impact of chronic diseases, such as coronary heart disease (CHD), cardiovascular disease (CVD), chronic kidney disease (CKD), diabetes, hypertension, and stroke will require integrated information from multiple systems.

We explored the possibility in this study to predict BA capable of reflecting the physiological or pathophysiological state in multiple organ systems by applying a multimodal image fusion AI model of retinal fundus, facial, and tongue images (16, 17). We hypothesize that tongue images may be a potential indicator for microbiome exposure and may reflect the state of oral and gastrointestinal track health (18, 19). We optimized this AI prediction model by exploiting image detail using a joint loss function to represent the progressive nature of aging and to tolerate minor errors in modeling. We trained and validated our AI model using fundus, facial, and tongue images from healthy participants

Author contributions: J.W., Y.G., J.Z., K.Z., and Y.Z. designed research; J.W., Y.G., J.Z., K.Z., and Y.Z. performed research; J.W., Y.G., J.Z., K.Z., and Y.Z. contributed new reagents/analytic tools; J.W., Y.G., F.W., S.Z., J.L., H.M., T.W., J.Z., D.B.-H., O.M., T.G., L.C., Y.L., Z.L., M.L., J.-k.Z., S.N., K.Z., and Y.Z. analyzed data; and J.W., Y.G., J.Z., K.Z., and Y.Z. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]J.W. and Y.G. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: niesheng0202@126.com, kang.zhang@gmail.com, or yongzhou78214@163.com.

Published January 8, 2024.

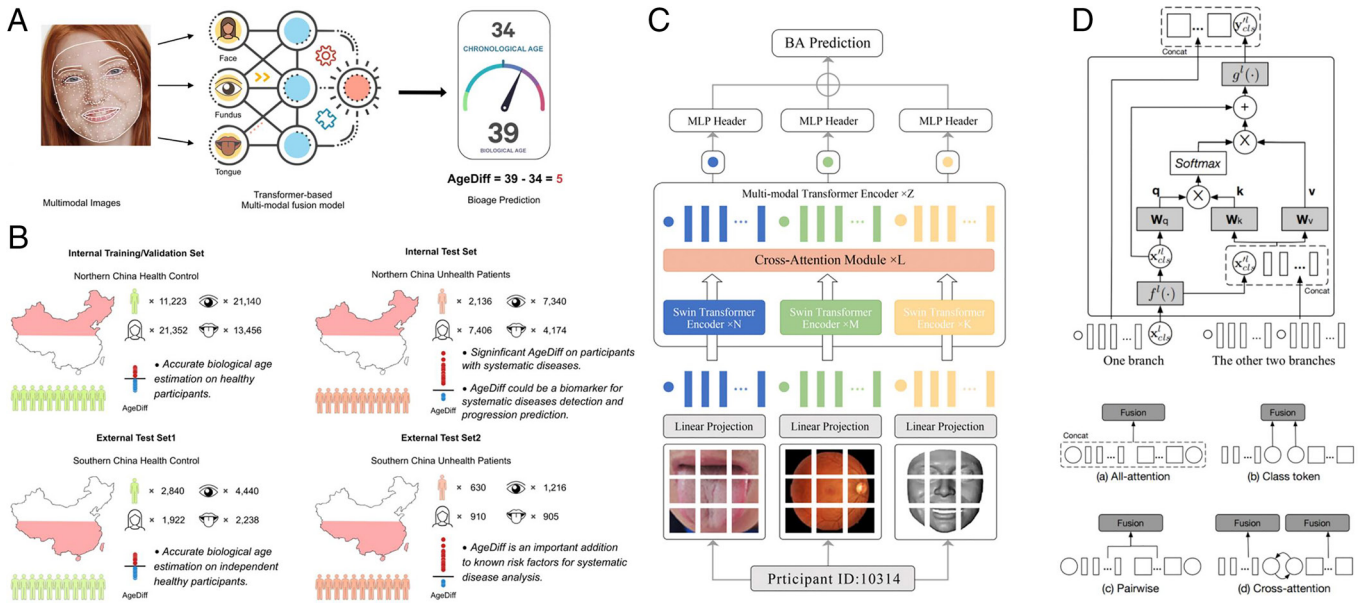

Fig. 1. (*A*) BA estimation using a combination of retinal images, tongue images, and facial images. (*B*) AI system was trained and validated on Northern China Healthy Control dataset, tested on Northern China Unhealthy dataset and Southern China datasets, demonstrating accurate BA estimation on healthy participants, significant AgeDiff in unhealthy participants, accurate AgeDiff-based ability for diagnosis and progression of systemic diseases, and strong potential to be an addition to known risk factors of systemic diseases. (*C*) AI system details: A Transformer-based architecture with a CAM for BA estimation. The networks are optimized using the loss between CA and BA with BP algorithm. (*D*, *Up*) a CAM which consists of a stack of Z MMT encoders. Each uses three different branches to process image tokens of different modalities and fuse the tokens at the end by an efficient module based on cross-attention of the CLS tokens (type d). (*Down*) Four types of multimodal fusion implementations. (*a*) All-attention fusion where all tokens are bundled together without considering any characteristic of tokens. (*b*) Class token fusion, where only CLS tokens are fused as it can be considered as global representation of one branch. (*c*) Pairwise fusion, where tokens at the corresponding spatial locations are fused together and CLS are fused separately. (*d*) Cross-attention, where CLS token from one branch and patch tokens from another branch are fused together.

and employed the model to estimate the impact of diseases and lifestyle factors on BA using images from participants with a number of chronic diseases and/or known risk factors for the development of chronic diseases. We showed that multimodal BA output is the closest to the true age in the healthy populations. BA is markedly increased in various diseases and unhealthy lifestyle habits and is a strong predictor of chronic diseases.

1. Results

1.1. Overview of the Model. An overview of our study incorporating the AI model is shown in Fig. 1. Our AI model is a Transformer-based architecture which incorporates a cross-attention module (CAM) for BA estimation using a combination of fundus, facial, and tongue images (Fig. 1*C*). The input images of three modalities

Table 1. Basic characteristics of the participants in the internal data set and the external data set

Cohorts	Normal	Any disease	CHD	CKD	CVD	Diabetes	Hypertension	Stroke
Northern China cohort								
Participants	11,223	2,136	321	935	354	523	1,686	57
Image	55,948	10,846	1,622	4,448	1,906	2,296	8,480	280
Face	21,352	7,406	610	1,706	702	11,004	3,280	104
Fundus	21,140	7,340	606	1,684	692	998	3,256	104
Tongue	13,456	4,174	406	1,058	158	694	1,944	72
Female (%)	5846 (52%)	983 (46%)	142 (44%)	452 (48%)	158 (44%)	249 (47%)	823 (49%)	25 (44%)
Age (y)	53.8 ± 11.3	56.7 ± 10.5	55.6 ± 10.9	57.2 ± 11.2	57.3 ± 10.8	56.6 ± 11.4	55.1 ± 10.8	57.4 ± 11.0
BMI (kg/m^2)	24.7 ± 2.3	25.0 ± 2.4	24.9 ± 2.2	24.8 ± 2.4	25.1 ± 2.2	25.0 ± 2.3	25.1 ± 2.3	25.2 ± 2.2
Smoking (%)	2531 (23%)	1329 (62%)	171 (53%)	379 (41%)	140 (40%)	325 (62%)	896 (53%)	3 (5%)
Drinking (%)	3716 (33%)	1405 (66%)	142 (44%)	514 (55%)	153 (43%)	318 (61%)	1045(62%)	9 (16%)
eGFR (mL/min per 1.73 m^2)	97.3 ± 22.5	101.5 ± 23.8	98.2 ± 22.9	103.2 ± 24.5	99.3 ± 22.0	100.5 ± 20.7	99.3 ± 23.1	98.3 ± 20.5
Blood glucose (mmol/L)	6.6 ± 2.3	7.1 ± 2.5	6.9 ± 2.8	7.0 ± 2.6	7.2 ± 2.3	7.1 ± 2.8	7.0 ± 2.9	6.9 ± 2.6
Southern China cohort								
Participants	2,840	630	43	–	36	124	510	36
Image	8,600	2,867	183	–	155	503	2,038	142
Face	1,922	910	55	–	40	156	614	45
Fundus	4,440	1,216	86	–	72	204	793	61

(Continued)

Table 1 (Continued)

Cohorts	Normal	Any disease	CHD	CKD	CVD	Diabetes	Hypertension	Stroke
Tongue	2,238	905	52	–	43	143	631	36
Female (%)	844 (29.7%)	98 (26.7%)	11 (34.8%)	–	7 (19.4%)	30 (24.2%)	134 (26.3%)	7 (19.4%)
Age (y)	49.8 ± 7.3	56.2 ± 9.8	55.4 ± 9.8	–	57.3 ± 10.8	57.2 ± 9.9	55.1 ± 10.8	57.4 ± 11.0
BMI (kg/m^2)	24.2 ± 3.6	24.9 ± 3.3	24.6 ± 3.2	–	24.8 ± 3.4	24.9 ± 3.7	25.4 ± 3.6	24.8 ± 3.4
Smoking (%)	752 (26.4%)	168 (26.8%)	7 (16.3%)	–	12 (33.3%)	34 (27.4%)	125 (24.5%)	12 (33.3%)
Drinking (%)	119 (42.0%)	264 (42.1%)	142 (44.0%)	–	13 (36.1%)	38 (30.6%)	209 (41.0%)	13 (36.1%)
Blood glucose (mmol/L)	5.6 ± 1.7	6.2 ± 2.1	5.7 ± 1.7	–	6.1 ± 2.0	7.6 ± 2.8	5.7 ± 1.4	6.1 ± 2.0

are first sent to three linear projection modules to construct the corresponding image tokens and classification tokens (CLS). These ViT-like tokens are regarded as the input of a multimodal Transformer (MMT) that contains Z-stack encoders with a CAM. Each CAM uses three branches to process image tokens of three modalities and fuses the tokens at the end based on the CLS tokens of CAM. A cross-attention fusion (Fig. 1 *D*, *Up*) strategy, which involves the CLS token of one modality and image tokens of the other two modalities, is used in our model and demonstrates advantages over other heuristic approaches (Fig. 1 *D*, *Bottom*). The outputs of the MMT encoders are linked to standard MLP headers for BA prediction. The whole architecture is optimized using the loss function between the CA and the predicted BA using a backpropagation algorithm.

1.2. Patient Characteristics. The general scheme of our study design and procedures are described in Fig. 1. The training dataset contains subjects in the northern China cohort who were followed longitudinally for regular health checks starting with a cross-sectional study. A total of 14,063 subjects consented to participate in our study. They were subjected to 3D face, tongue, and retinal scanning, and relevant metadata were extracted from their medical

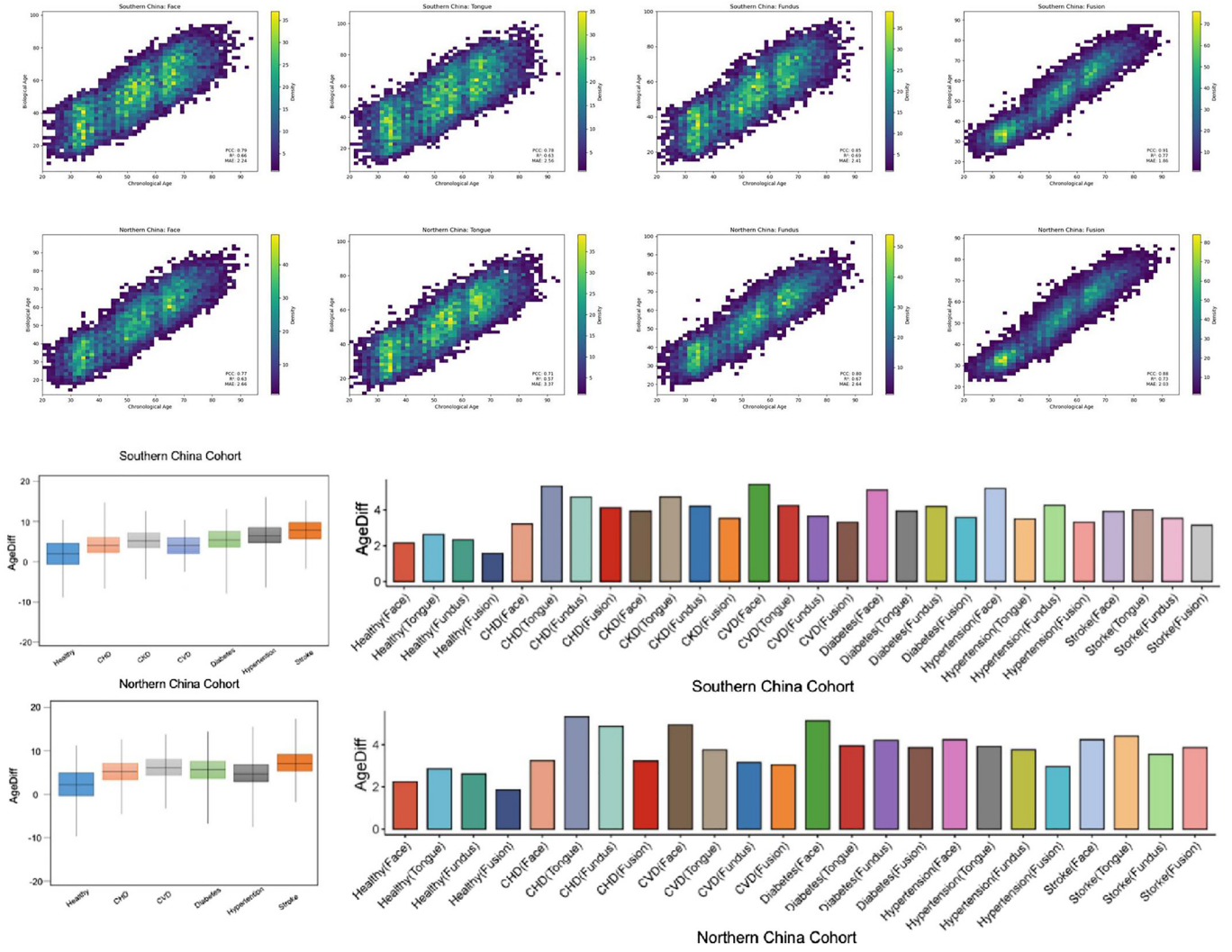

Fig. 2. Impact of chronic diseases and environmental factors on BA in both the internal and external cohorts. Correlation analysis of the predicted BA versus CA generated using the multimodal fusion architecture on the internal test set and external test set.

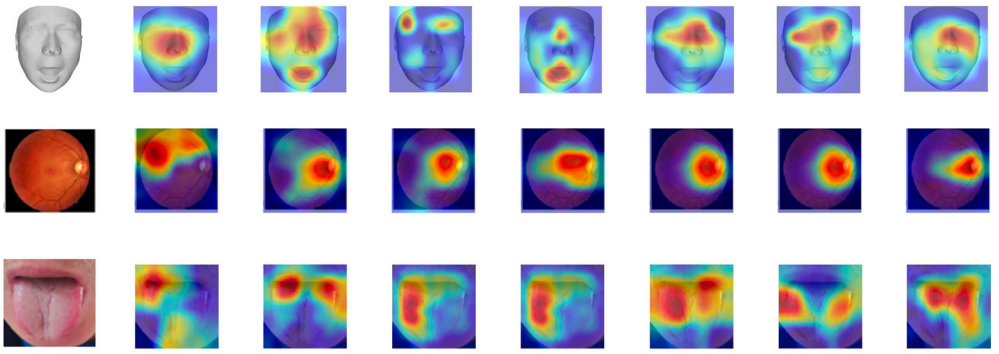

Fig. 3. Example of Grad-CAM++ results on three-modality inputs of one participant on internal training set at (100, 250, 300, 350, 400, 450, and 500) training epochs. The saliency maps gradually provide visual clues in the training process where the network is optimized with the loss between BA and CA, showing evolution of different attended regions and a final localized region at last. The final ROI was the most meaningful one for BA prediction. The ROI for retinal images was on the vessel dense areas which may suggest the importance of blood vessels in BA-related phenotypes. The ROIs for the face and tongue were on the eye and the tongue fur.

records. Blood was drawn after fasting followed by medical follow-up. The metadata included demographic information, lifestyle (including smoking and alcohol use), and outcomes from routine physical examinations and clinical laboratory assays (Fig. 1 and Table 1). All participants from the discovery cohort were split into mutually exclusive sets for training, tuning, and internal validation of the AI algorithm at an 80%:10%:10% ratio. The southern China cohort of 2,766 subjects serves as an independent validation cohort.

1.3. BA Estimation by Facial, Tongue, and Retinal Images. We applied a multimodal image fusion approach, using fundus, tongue, and facial images, in our AI model to estimate BA (Fig. 2). We trained our AI model using images from healthy participants to predict BA. We determined the accuracy of the AI model predicted BA by its difference from the CA of the corresponding participant using healthy participants. The scatter plots of BA predictions from the test sets in internal and external cohorts are shown in Fig. 2. In both cohorts, BA predictions using the multimodal image fusion approach produced a better correlation with the CA [Pearson's correlation coefficient (PCC) of 0.91 in the internal cohort and PCC of 0.88 in the external cohort]. The mean absolute error (MAE) as well as the Coefficient of determination (R^2) were also improved. Using Grad-CAM++ as an interpretation for our AI findings, we observed that our multimodal fusion AI model paid more attention to regions near the lip and center in tongue image, vascular-density region in retinal fundus image, and eye region in facial image (Fig. 3). Our data therefore indicate that our multimodal image fusion AI model was able to accurately predict BA, and was superior to BA prediction using either of the three image modalities alone.

We next used our multimodal image fusion AI model to evaluate the impact of chronic diseases and environmental factors on BA in both the internal and external cohorts (Fig. 2). We predicted the BA of each subject and evaluated the AgeDiff, as above. We plotted the mean AgeDiff (Fig. 4) and found that in individuals with chronic diseases, the predicted BA was higher than the CA when compared to the age difference in healthy participants by AgeDiff of 3.16 y in CHD (95% CI, 2.67 to 3.62; *P*-value < 0.001), 3.85 y in CKD (95% CI, 3.43 to 4.35; *P*-value < 0.001), 4.51 y in CVD (95% CI, 3.77 to 5.23; *P*-value < 0.001), 3.94 y in diabetes (95% CI, 3.58 to 4.43; *P*-value < 0.001), 4.06 y in hypertension (95% CI, 3.74 to 4.33; *P*-value < 0.001), and 4.94 y in stroke (95% CI, 4.13 to 5.48; *P*-value < 0.001) (Fig. 2). Interestingly, we also observed an AgeDiff of 5.43 y in smokers (95% CI, 4.56 to 6.13; *P*-value < 0.001), AgeDiff of 3.62 y in drinkers (95% CI, 3.45 to 4.16; *P*-value < 0.001), and an AgeDiff of 4.36 y in obese participants (BMI > 27, 95% CI, 3.71 to 4.82; *P*-value < 0.001).

1.4. Prediction of Chronic Diseases Risks using AgeDiff. We categorized the difference between BA and CA into four equal quartiles in an attempt to stratify our analyses on the basis of the

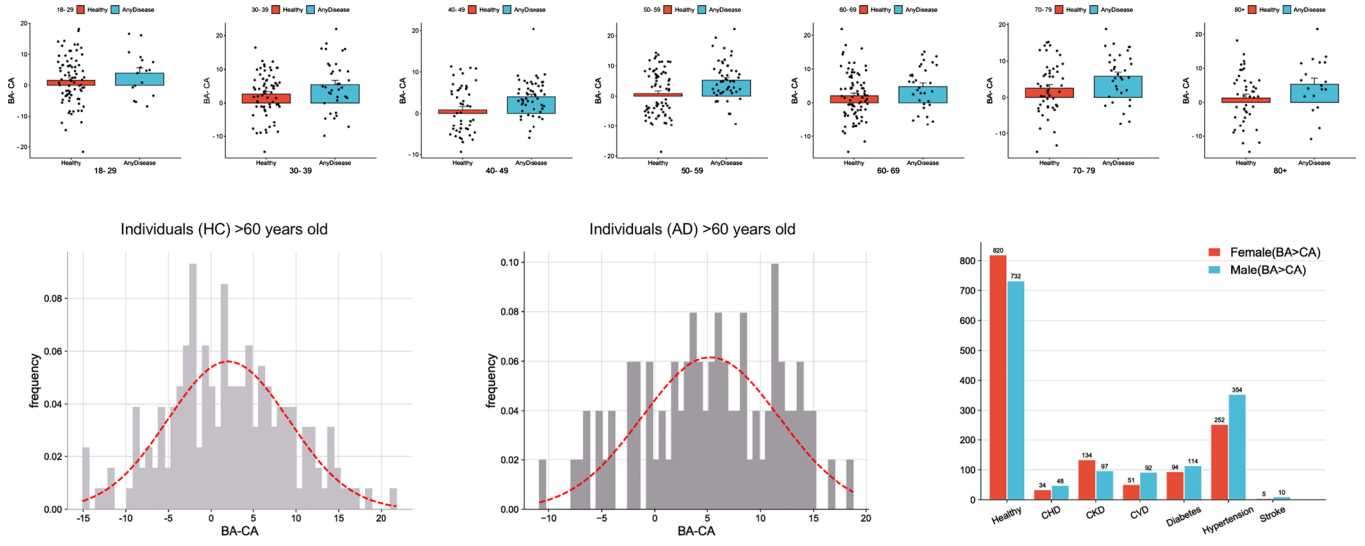

Fig. 4. Comparison of AgeDiff for individuals healthy or of any disease for each decade of life. Statistics comparison of AgeDiff of the two groups using Student's *t* test (*P* < 0.05 for all groups, HC: healthy control; AD: any disease). Calculation of AgeDiff is based on the difference between an individual's predicted BA and actual CA. Sex distribution for individuals of increased age (BA > CA) in each group.

Table 2. Association between the AgeDiff with the incident of six common chronic systematic diseases

Cohorts		Any disease		CHD		CKD		CVD		Diabetes		Hypertension		Stroke	
Internal test set AgeDiff															
All participants		HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value
Mean (SD) (4.56)	2.32	1.5 (1.31-2.11)	0.015	1.9 (1.70-2.21)	0.031	1.1 (1.02-1.32)	0.018	1.4 (1.12-1.69)	0.023	1.5 (1.26-1.77)	0.042	2.0 (1.74-2.14)	0.028	1.3 (1.09-1.44)	0.041
Quartile 1	−7.23 (3.05)	1 [Reference]	–	1 [Reference]	–	1 [Reference]	–	1 [Reference]	–	1 [Reference]	–	1 [Reference]	–	1 [Reference]	–
Quartile 2	−2.59 (1.31)	1.34 (1.13-1.51)	0.116	1.72 (1.43-1.91)	0.108	1.32 (1.09-1.46)	0.043	1.23 (1.09-1.47)	0.192	1.33 (1.13-1.71)	0.113	1.45 (1.15-1.72)	0.071	1.62 (1.23-1.81)	0.194
Quartile 3	4.18 (1.78)	2.15 (1.74-2.53)	0.024	1.76 (1.24-2.23)	0.043	2.57 (1.91-3.62)	0.012	2.96 (1.94-3.55)	0.041	2.36 (1.42-3.31)	0.035	2.61 (1.94-3.60)	0.043	2.35 (1.65-3.31)	0.038
Quartile 4	8.25 (2.70)	5.72 (4.59-6.11)	0.007	5.04 (4.29-6.42)	0.022	5.25 (4.41-6.06)	0.005	5.16 (4.34-5.74)	0.010	5.61 (4.52-6.35)	0.026	5.78 (4.71-7.21)	0.021	4.67 (4.10-5.53)	0.018
External test set AgeDiff															
All participants		HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	–	–	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value
Mean (SD) (4.13)	2.07	1.4 (1.10-1.62)	0.031	1.6 (1.18-1.73)	0.071	–	–	1.5 (1.31-1.77)	0.013	1.3 (1.12-1.74)	0.038	1.7 (1.54-2.05)	0.037	1.1 (1.04-1.32)	0.025
Quartile 1	−8.12 (3.43)	1 [Reference]	–	1 [Reference]	–	–	–	1 [Reference]	–	1 [Reference]	–	1 [Reference]	–	1 [Reference]	–
Quartile 2	−4.29 (1.85)	1.55 (1.22-1.74)	0.046	1.72 (1.43-1.91)	0.112	–	–	1.43 (1.12-1.63)	0.132	1.33 (1.13-1.71)	0.103	1.45 (1.15-1.72)	0.043	1.54 (1.28-1.79)	0.033
Quartile 3	2.13 (1.72)	1.87 (1.44-2.15)	0.015	3.06 (2.39-3.63)	0.029	–	–	2.43 (1.74-3.10)	0.021	2.28 (1.73-3.04)	0.025	3.41 (2.13-3.78)	0.013	2.65 (1.85-3.21)	0.011
Quartile 4	6.35 (2.32)	4.67 (3.34-6.52)	0.002	5.53 (3.81-6.79)	0.004	–	–	4.16 (3.64-5.29)	0.009	4.61 (4.05-6.24)	0.016	5.68 (4.71-7.21)	0.008	5.67 (4.10-6.56)	0.005

The first quartile (Q1) is defined as the set of data between the smallest value and the 25th retinal age gap. The second quartile (Q2) is the set of data between the 25th and median value. The third quartile (Q3) is set of data between the median value and the 75th retinal age gap. The fourth quartile (Q4) is defined as the set of data between the 75th and the maximum of the retinal age gap.

BA difference. We evaluated the hazard ratio (HR) of developing each of the chronic diseases in each of these quartiles. The results of our analyses are shown in Table 2. Overall, changes in the AgeDiff were associated with the development of any types of the six chronic diseases in the internal cohort (HR = 1.5, 95% CI = 1.70 to 2.11, $P = 0.015$) and external cohort (HR = 1.4, 95% CI = 1.10 to 1.63, $P = 0.031$). For the individual chronic diseases evaluated, changes in BA were associated with an increased HR for developing each of the diseases analyzed in the internal cohort (hypertension, CHD, diabetes, CVD, stroke, and CKD) and external cohort (hypertension, CHD, CVD, diabetes, and stroke). Within the different quartiles, there was an overall trend for increasing HR for developing each of the chronic diseases with successive quartiles. In both cohorts, quartile 4 was significantly associated with higher HR for developing chronic diseases, while there were no significant associations in quartiles 1 and 2. In the internal cohort, patients in quartile 3 were significantly associated with higher HR for diabetes and stroke, while external participants in quartile 3 were significantly associated with CVD. The association between BA difference and HR for developing these common chronic diseases remained statistically significant even following the removal of participants who were diagnosed with these diseases within 1 y (Table 2).

We subsequently evaluated the utility of our multimodal image fusion model on AgeDiff to predict the 5-y risks of developing CHD, CVD, CKD, stroke, hypertension, and diabetes and compared these predictions to standard approaches using established risk factors. Among these risk factors, we find that the body mass index and diastolic blood pressure to have the largest impact on predicted BA, using SHAP (SHapley Additive exPlanations) analyses (Fig. 5). The receiver operator characteristic (ROC) curves and precision–recall curves (PRCs) for prediction of chronic disease development are shown in Fig. 6. We found that the predictive value of the AgeDiff for chronic diseases, evaluated using area under the both ROC and PRC measurements, was consistently higher, relative to predictions using known risk factors. Importantly, we found that combination of BA difference with risk factors improved the AUC under both ROC and PRC, indicating that BA prediction can be used in conjunction with existing risk factors to identify individuals at risk of developing chronic disease.

1.5. Incidence Prediction of Chronic Diseases using AgeDiff. Our results so far demonstrate that BA difference can be used to predict the risk of developing chronic diseases. Next, we evaluated whether our BA prediction model can be used to predict the disease onset. The performance of incidence prediction for different chronic diseases using BA difference under the Cox proportional hazards (CPH) model is summarized in Table 3. Similar to our observations above, combination of the BA difference and the risk factor–based model provided an improved C-index for the incidence detection of chronic diseases. When testing on another independent external cohort, similar results were observed. The above results show that BA, as an important biomarker, could be used to assist existing factors for disease prognosis.

We used the Kaplan–Meier method to stratify healthy individuals at the baseline into two risk groups (low or high risk) for developing

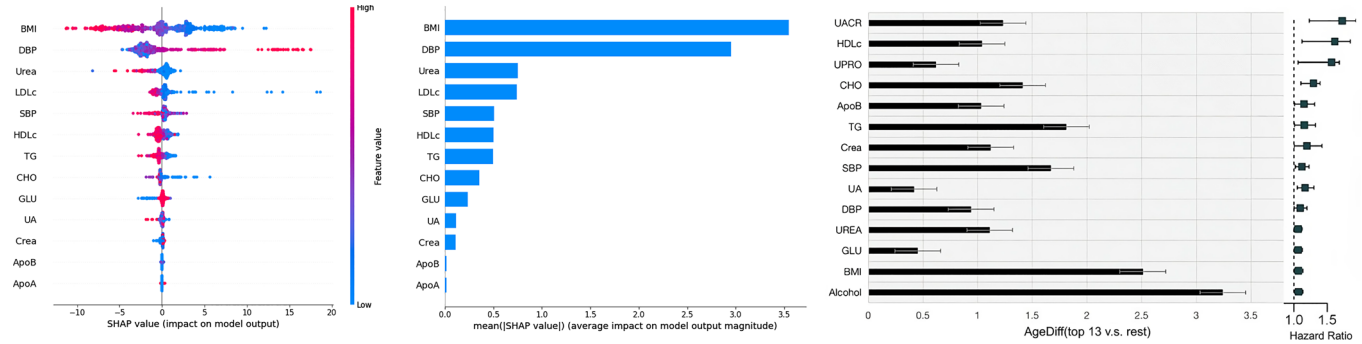

Fig. 5. Top 13 variants in terms of attributable AgeDiff using the SHAP method (*Left* and *Middle*). Top 13 variants in terms of attributable AgeDiff and HRs for any disease (*Right*). Estimates are based on the internal test set. Error bars denote 95% CI.

chronic diseases. The incidence of the different diseases stratified by risk groups of the BA difference model is shown in Fig. 7. For the Kaplan–Meier curves and log-rank tests, thresholds for the high-risk and low-risk groups were based on the upper and lower quartiles of the predicted risk scores from the combined models in the training cohort. We then tested our approach on the test cohort and found statistically significant separations of the low-risk and high-risk groups (Fig. 7). Our data therefore indicate that our multimodal image fusion AI model was able to identify at-risk patients for chronic diseases and predict chronic disease incidence.

1.6. Relations between AgeDiff and Other Risk Factors. According to previous studies (20, 21), the six chronic diseases included in this study have been associated with various risk factors, among

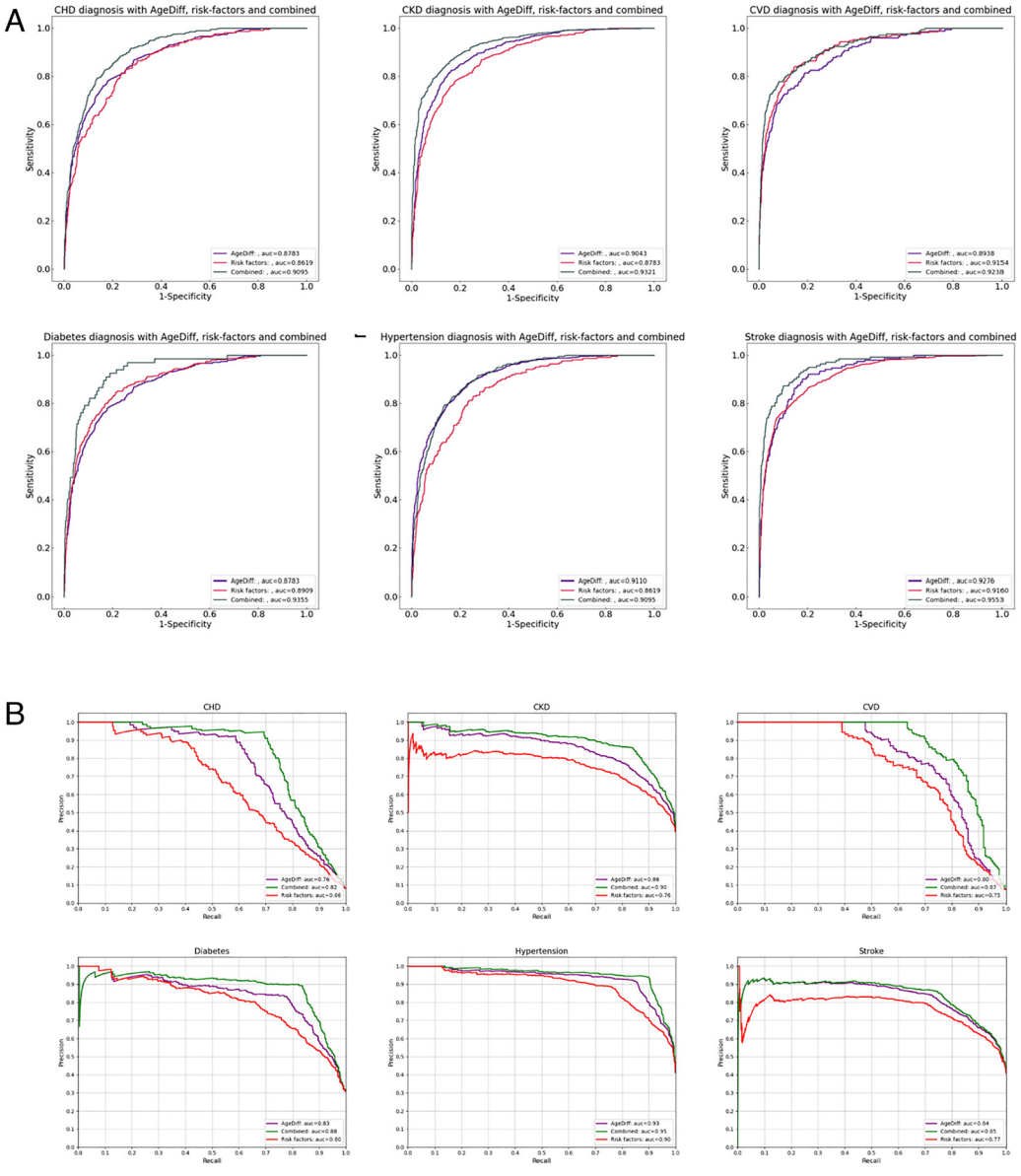

Fig. 6. Performance of the AI models in the identification of six common chronic systematic diseases on the internal test set using the risk-factor-only model, the multimodal fusion model, and the combined model. (*A*) The ROC and the AUC score under ROC for each disease. (*B*) The PRC and the AUC score under PRC for each disease.

Table 3. Performance of progression prediction model to six common chronic systematic diseases event based on the risk-factor-only model, and the combined model (including multi-modal images and risk-factors) on the internal and external test sets

	Progression prediction models	C-index on internal test set	C-index on external test set
CHD	Risk-factor-based model	0.775 (95% CI: 0.719-0.850)	0.813 (95% CI: 0.726-0.853)
	BA-based model	0.825 (95% CI: 0.726-0.894)	0.848 (95% CI: 0.751-0.896)
	Combined model	0.853 (95% CI: 0.812-0.913)	0.872 (95% CI: 0.830-0.925)
CKD	Risk-factor-based model	0.828 (95% CI: 0.753-0.916)	–
	BA-based model	0.813 (95% CI: 0.734-0.904)	–
	Combined model	0.865 (95% CI: 0.768-0.935)	–
CVD	Risk-factor-based model	0.806 (95% CI: 0.731-0.901)	0.803 (95% CI: 0.753-0.861)
	BA-based model	0.819 (95% CI: 0.758-0.896)	0.841 (95% CI: 0.788-0.899)
	Combined model	0.856 (95% CI: 0.788-0.924)	0.857 (95% CI: 0.801-0.905)
Diabetes	Risk-factor-based model	0.868 (95% CI: 0.761-0.915)	0.803 (95% CI: 0.751-0.882)
	BA-based model	0.867 (95% CI: 0.772-0.927)	0.857 (95% CI: 0.781-0.905)
	Combined model	0.903 (95% CI: 0.824-0.942)	0.872 (95% CI: 0.814-0.933)
Hypertension	Risk-factor-based model	0.813 (95% CI: 0.712-0.890)	0.803 (95% CI: 0.743-0.866)
	BA-based model	0.826 (95% CI: 0.735-0.912)	0.826 (95% CI: 0.778-0.894)
	Combined model	0.874 (95% CI: 0.788-0.939)	0.854 (95% CI: 0.792-0.915)
Stroke	Risk-factor-based model	0.872 (95% CI: 0.773-0.920)	0.810 (95% CI: 0.753-0.864)
	BA-based model	0.861 (95% CI: 0.756-0.917)	0.834 (95% CI: 0.796-0.894)
	Combined model	0.895 (95% CI: 0.842-0.935)	0.876 (95% CI: 0.821-0.921)

Concordance index (C-index) for right-censored data and 95% CI measure the model performance by comparing the progression information (disease labels and progression days) with predicted risk scores. A larger C-index correlates with better progression prediction performance. CI, confidence interval.

which we collected 61 covariates. We first conducted univariate and multivariate survival analyses using CPH methods (likelihood ratio), including AgeDiff and other prognostic factors, in addition to the scores generated from the six chronic diseases. As Table 4 shows, under both univariate and multivariate analysis, AgeDiff is proved to be a significant factor for developing chronic diseases. We further examined the relations between AgeDiff and other risk factors, investigating the most relevant risk factors related to AgeDiff. To this end, we built a lightgbm (22) which is a gradient boosting framework that uses tree based learning algorithms for mapping 61 factors to AgeDiff. Fig. 5 shows top 13 D attributable variants to AgeDiff using the SHAP method (23) (*Left* and *Middle*). We also illustrated top 13 variants in terms of attributable AgeDiff and HRs for any diseases (*Right*). These results provide explainable contributors to AgeDiff and chronic diseases.

2. Discussion

In this paper, we proposed a multimodal fusion framework that incorporates facial, tongue, and retinal image detail enhancement and a joint loss function for BA prediction. Our model was validated using an independent dataset and demonstrated robustness, the ability to reflect the progressive nature of aging, and improved predictive accuracy compared to the recently reported approaches for BA prediction using retinal age (8). While previous studies have demonstrated facial or retinal age to be a biomarker of aging (8, 10, 11, 24), our study expanded this potential by combining and integrating retinal, tongue, and facial images to gain a more complete portrait of BA. Our AI model achieved comparable BA prediction on retinal age to previous studies (around 2.5 y versus CA). However, when combined with facial

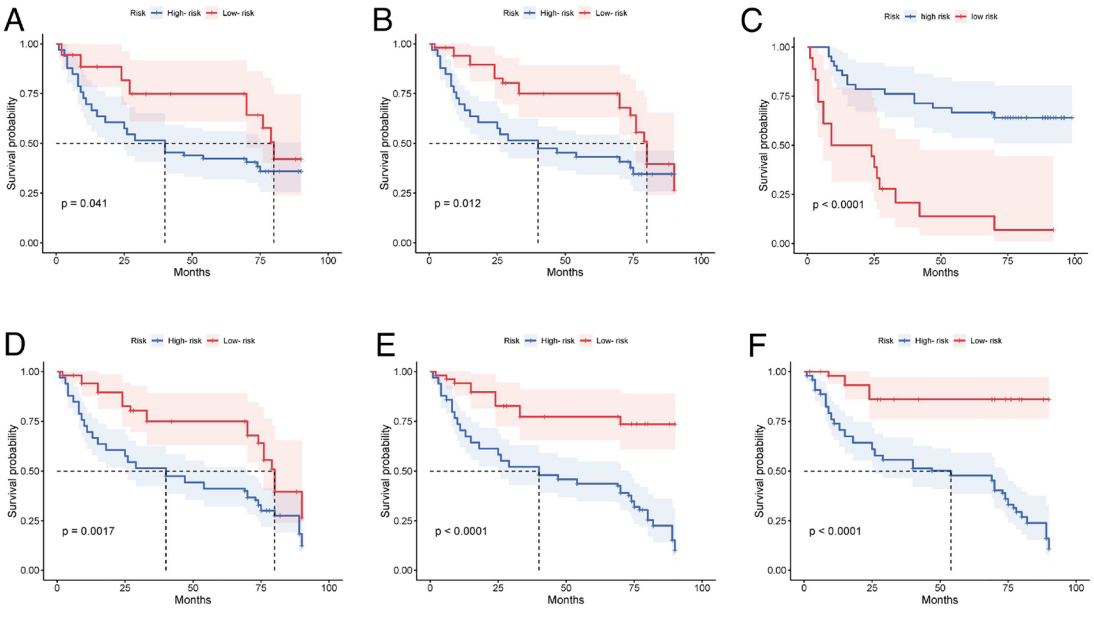

Fig. 7. Kaplan–Meier plots for the prediction of six common chronic systematic diseases (*A*: CHD, *B*: CKD, *C*: CVD, *D*: Diabetes, *E*: Hypertension, *F*: Stroke) on the internal test set. The *y* axis is the survival probability, measuring the probability of not progressing to a disease outcome. The *x* axis is the time in months. Survival curves in different colors represent the high-risk and low-risk subgroups stratified by the upper quartiles in the tuning dataset. Shaded areas are 95% CI.

Table 4. Predicted incidence rates of six common chronic systematic diseases (per 1,000 person-years) for the internal longitudinal test set and for the external longitudinal test set, stratified by risk level

Disease	Subset	Participants	Events	Incident rate	Univariate analysis		Multivariate analysis	
					HR (95%CI)	*P* value	HR (95% CI)	*P* value
Prognostic analysis on internal longitudinal test set								
CHD	Low risk	1,029	31	3.0 (0.6, 9.5)	Reference	NA	Reference	NA
	High risk	1,063	94	8.6 (3.8, 17.8)	5.7 (2.4, 8.0)	<0.001	2.2 (0.9, 5.3)	<0.001
CKD	Low risk	1,854	102	5.5 (1.3, 9.6)	Reference	NA	Reference	NA
	High risk	1,771	280	15.8 (4.9, 23.4)	9.2 (3.3, 14.5)	<0.001	6.4 (3.8, 9.6)	<0.001
CVD	Low risk	1,317	25	1.9 (0.1, 4.1)	Reference	NA	Reference	NA
	High risk	1,392	77	5.5 (3.8, 8.5)	3.1 (0.7, 5.6)	<0.001	1.7 (0.3, 4.2)	<0.001
Diabetes	Low risk	1,648	55	3.3 (0.5, 6.7)	Reference	NA	Reference	NA
	High risk	1,715	110	6.4 (5.8, 11.5)	2.6 (1.1, 4.6)	<0.001	2.1 (1.3, 3.2)	<0.001
Hypertension	Low risk	2,683	157	6.0 (2.1, 9.4)	Reference	NA	Reference	NA
	High risk	3,297	316	9.6 (5.8, 15.5)	5.3 (2.4, 8.0)	<0.001	3.7 (1.6, 5.2)	<0.001
Stroke	Low risk	1,492	11	0.7 (0.0, 2.7)	Reference	NA	Reference	NA
	High risk	1,384	5	0.3 (0.0, 2.4)	2.3 (1.8, 2.8)	<0.001	1.9 (1.4, 2.4)	<0.001
Prognostic analysis on external longitudinal test set								
CHD	Low risk	169	8	2.3 (1.4, 3.5)	Reference	NA	Reference	NA
	High risk	177	13	5.3 (1.9, 4.7)	4.7 (2.1, 7.5)	<0.001	3.2 (1.9, 5.1)	<0.001
CVD	Low risk	125	5	1.7 (1.1, 2.7)	Reference	NA	Reference	NA
	High risk	332	16	4.5 (3.8, 8.5)	3.3 (1.7, 5.2)	<0.001	2.7 (0.9, 4.8)	<0.001
Diabetes	Low risk	204	32	4.5 (1.5, 7.9)	Reference	NA	Reference	NA
	High risk	425	45	7.4 (3.8, 12.2)	4.3 (1.4, 5.2)	<0.001	4.1 (1.2, 4.7)	<0.001
Hypertension	Low risk	191	72	6.0 (4.1, 8.2)	Reference	NA	Reference	NA
	High risk	367	151	11.6 (5.1, 16.2)	6.4 (3.3, 8.9)	<0.001	4.3 (2.6, 5.9)	<0.001
Stroke	Low risk	152	4	1.1 (0.1, 2.9)	Reference	NA	Reference	NA
	High risk	324	8	2.4 (0.1, 2.5)	2.9 (1.4, 4.1)	<0.001	2.1 (1.5, 2.7)	<0.001

and tongue images, our multimodal AI achieved BA predictions within 2 y for healthy individuals. The proposed BA estimation method is more accurate than other phenotypic BA prediction studies by 2-5 y (6, 25–27). It is superior to established BA prediction models such as DNA methylation clocks (5, 28), transcriptome aging clocks (27, 29), and blood profiles (30, 31). Our AI model also shows statistically significant differences in BA between healthy and diseased subjects, indicating that the impact of diseases in BA and the potential of BA-based AgeDiff as an effective biomarker of aging and age-related disease research. Our study showed a link between accelerated BA and risk of chronic diseases such as CHD, CVD, CKD, stroke, hypertension, and diabetes.

Prediction of tissue and organ age is currently exemplified by retinal age, which is able to correlate between retinal neuronal and vascular changes and age-related brain diseases (12, 32). This raises the possibility of using retinal age as a surrogate measure of brain and vascular BA. The retina and cerebrum do share high similarities in microvasculature (33) and aging outcomes, such as the accumulation of mitochondria oxidative stress (34). However, we argue that BA predictions based on single organ systems, while useful to offer insight into system-specific diseases, do not offer a sufficiently accurate prediction of the overall physiological or pathophysiological state of the individual. Facial and tongue images may therefore add other dimensions to accurately estimate BA. Several population-based studies (10, 35, 36) have shown that aging concomitantly alters the retina, brain, skin, and the gastrointestinal tract. Indeed, it is possible that tongue health may offer a window into gastrointestinal tract status and also microbiome exposure (18, 19). Facial images may offer an assessment of direct sun and air exposure. These links will require further investigation, and will undoubtedly uncover interesting, and important relationships between chronic diseases and tongue and facial features. Nevertheless, our results showing that the predicted BA using fundus images can be improved by incorporating facial and tongue images supports our argument that tongue and facial images, when combined with AI, may offer insights into an individual's overall physiology.

Our study shed light on the potential for a multimodal image–based AI prediction to become a large-scale screening tool for individuals at high risk for various chronic diseases. The BA predictions based on our model offer unique advantages of detecting the risk, as well as prognosis, of a range of diseases through a fast, noninvasive, and economical method. Additionally, these predictions can be made even more accessible by incorporating smartphone-based teleophthalmology and facial and tongue imaging assessment (37). There have been ethical and privacy concerns with using facial images for BA prediction. However, we believe these concerns will be somewhat mitigated with our fusion approach since the facial images are combined with fundus and retinal images for analyses. This potential will be approached with the increasing ability of transformer-based AI (38) and large language models (24). In conclusion, our study revealed the potential utility of using multimodal images to predict BA, which can be used to identify individuals at risk of developing chronic diseases and to intervene so the disease risks can be reduced.

3. Methods

3.1. Image Datasets and Patient Characteristics. The 3D facial, tongue, and retinal images were collected from the study cohorts of the China Bioage Investigation Consortium, which consists of the following participants: the northern China cohort which was used for the model training and the southern China cohort, which is used for an independent validation. The northern China cohort is from the China suboptimal health cohort study (COACS) in Tangshan City, Hebei Province, China. The southern China cohort is from the Nanfang Hospital in Guangzhou, Guangdong Province, Zhuhai People's Hospital. Institutional Review Board approvals were obtained in COACS, Nanfang Hospital and Zhuhai People's Hospital, and all participating subjects signed an informed consent form.

The COACS is a community-based, prospective study, to investigate how suboptimal health status contributes to the incidence of noncommunicable chronic diseases in Chinese adults (39). This COACS study is a cross-sectional survey. The participants were recruited from Tangshan city, which is a large, modern industrial city adjacent to two megacities: Beijing and Tianjin. All participants underwent clinical, laboratory, and environmental exposure measurements aimed at identifying clinical, biological, environmental, and genetic factors associated with suboptimal health. We have elected to use this cohort for our study because it has the balance of healthy subjects and those with metabolic diseases, medical records were relatively complete, and previous electronic medical records were available for assessment if needed. The southern China cohort is also a community-based, annual health-check prospective study with a similar study design.

The northern China developmental cohort and the southern China validation external consisted of patients with demographic information and clinical parameters from their electronic medical records. If they consented to this study, they were subjected to 3D face, tongue, and retinal scanning, fasting blood draws, and the use of medical record data. 3D facial images were captured using 3dMDface camera systems (www.3dmd.com) with the study beginning in their annual visit in 2018 to 2022. Applying standard facial and retinal image acquisition protocols, participants were asked to close their mouths and hold their faces with a neutral expression for the capture of the digital facial stereophotogrammetry. 3D images in wavefront.obj file format with point clouds and corresponding texture images were used for further analysis. For each consenting subject, demographic, routine physical examination, and clinical laboratory were obtained. Demographic and clinical data for all the study participants are summarized in Table 1.

3.2. Data Preprocessing. Our multimodal fusion architecture received three inputs, the integrated tongue, retinal fundus, and facial images. The size of each image was resized to 256×256. Tongue and facial images included learnable parameters that were optimized along with our multimodal fusion architecture.

Tongue images in this study were captured using standard settings on an iPhone X. Samples which were corrupt, vague, or those with strong illumination were excluded from the analysis. Nontongue elements, such as the face, teeth, lips, and neck, were removed using a preprocessing segmentation step. This involved coarse segmentation and fine segmentation to obtain pixel-level tongue contour, which is superior to rectangular ROI detection approaches. Rectangular ROI was produced in the coarse step, which formed the input for the fine segmentation. We used decorrelation stretch algorithm (28) equipped with the OSTU method (40) to attain an edge map. The tongue contour obtained from the improved maximal similarity-based region merging method (41) was then combined with the edge map to generate a weight map of the equal size to the original tongue image. Finally, the edge-based method fast marching (42) was implemented on the weight map to compute the final tongue contour. Once a precise tongue contour was obtained, it was converted into three spaces using three learnable modules (ColorNet, TextureNet, and GeometryNet), and leveraged their integrated image as the input of our multimodal fusion architecture. ColorNet consists of three multilayer perceptrons (MLPs) which take as input the conversion output from the original RGB contour using standard RGB-CIE mapping. TextureNet consists of three MLPs which take as input the RGB channels. GeometryNet consists of a three-layer MLP that receive the gray version of the contour and a linear embedding that takes as input the key landmark points (43).

The retinal fundus images were captured using standard fundus cameras, including Topcon TRC-NW6 (Topcon), Zeiss Visucam 224 (Carl Zeiss Meditec AG), Canon CR6-45NM (Canon), and KOWA Nonmyd α-DIII (Kowa). All fundus images were deidentified. For screening and grading retinal fundus images, a hierarchical two-tier grading process was performed by 10 phase I and five phase II graders. Phase I graders consisted of individuals trained by ophthalmologists and evaluated to attain at least 95% accuracy determined by a quiz consisting of 1,000 fundus images of various retinal diseases. Phase II graders consisted of ophthalmologists who individually reviewed every image classified by phase I graders. To check consistency among phase II graders, 20% of images were randomly selected and reviewed by three senior retinal specialists. The second tier of five ophthalmologists independently read and verified the true labels for each image. To account for disagreement, the evaluation test set was also checked by expert consensus.

The 3D facial stereophotogrammetry images were captured with standard acquisition protocols, where participants were asked to close their mouths and hold their faces with a neutral expression. Each 3D facial image included a 3D mesh and a corresponding texture image, extracted for each point and constructed into an integrated facial image as the input of multimodal fusion architecture. The texture features were expressed with the color of each point in a 3D facial image mapped through captured 2D texture images and texture coordinates to describe the photometric and color attributes of the face. Geometry features include global geometry features and local geometry features. Global features included the sizes of the

Table 5. Ablation study on BA prediction and AgeDiff-based diagnostic predictions using various configurations of input modality

Cohort	One			Two			Three
Northern/Southern China	Fundus	Face	Tongue	Fundus face	Fundus tongue	Face tongue	Fundus face tongue
MAE between BA and CA							
Normal group	3.32	4.10	5.65	3.65	3.26	2.64	1.94
Abnormal group (six disorders)	5.12	5.27	6.22	4.21	5.67	5.09	3.52
AUC-ROC performance using AgeDiff							
CHD	0.84	0.85	0.80	0.79	0.83	0.83	0.88
CKD	0.87	0.82	0.83	0.81	0.83	0.85	0.90
CVD	0.81	0.85	0.84	0.82	0.84	0.82	0.89
Diabetes	0.82	0.86	0.81	0.86	0.83	0.81	0.88
Hypertension	0.83	0.82	0.79	0.84	0.81	0.80	0.91
Stroke	0.85	0.84	0.76	0.83	0.87	0.82	0.92

Same architecture and implementation strategy was used with suppression of each modality in order to examine the contributions of each of the three modalities. The results on the test sets of two cohorts demonstrate that fundus images are the most important factor for an accurate prediction of BA (face and tongue weighted the second and the third in term of contribution) as well as analysis of related diseases, and a multimodal fusion of three modalities obtained the best performance.

Table 6. BA estimation and AgeDiff-based diagnostic predictions using our model and four popular deep learning models including 2 CNN-based ones (VGG-16 and ResNet-34) and 2 transformer-based ones (VIT-base-patch16-224 and Swin-tiny-patch4-window7-224)

Northern/Southern China cohort	Ours	VGG-16	ResNet-34	VIT	Swin-T-V2
MAE between BA and CA					
Normal group	1.94	5.32	6.21	5.58	4.73
Abnormal group (six disorders)	3.52	7.17	7.45	6.37	6.28
AUC-ROC performance using AgeDiff					
CHD	0.88	0.82	0.81	0.79	0.81
CKD	0.90	0.84	0.77	0.81	0.82
CVD	0.89	0.81	0.82	0.80	0.77
Diabetes	0.88	0.80	0.82	0.86	0.80
Hypertension	0.91	0.83	0.79	0.74	0.83
Stroke	0.92	0.87	0.76	0.82	0.81

These 4 models were pretrained on ImageNet and fine-tuned on the validation sets with late fusion strategy instead of CAM used in our implementations. Results were on the test sets of our two cohorts.

whole mesh and feature map of each component with three channels comprising the 3D coordinates of each point. Local features included shape depressions and prominences that were quantified by normal vectors and surface curvatures at each point in the mesh. We calculated Gaussian curvature and mean curvature of curvature for each point. Finally, global and local geometry maps as well as texture maps were integrated to generate a facial image.

3.3. Multimodal Fusion Architecture Settings. The number of MMT encoders K was set to 3. The numbers of Swin-Transformer (44) encoders for each modality were set to $M = 4$, $N = 4$, and $K = 5$. The number of Swin-Transformer encoders of CAMs in one MMT encoder was set to $L = 3$. The expanding ratio of feed-forward network in the Swin-Transformer encoder was set to 4. The number of headers were the same and set to 3 for three branches. Each of the two hidden

Table 7. Univariate and multivariate survival analyses of six common chronic systematic diseases conducted using CPH methods (likelihood ratio test)

		Univariate analysis		Multivariate analysis			Univariate analysis		Multivariate analysis	
Covariates	Disease	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value	Disease	HR (95% CI)	*P*-value	HR (95% CI)	P-value
	CHD					Diabetes				
Sex		0.94 (0.71-0.99)	<0.001	0.91 (0.73-1.07)	<0.001		0.65 (0.56-0.76)	<0.001	1.01 (0.82-1.24)	<0.001
BMI		1.24 (1.06-1.31)	<0.001	1.14 (1.04-1.21)	<0.001		1.16 (1.14-1.18)	<0.001	1.08 (1.04-1.11)	<0.001
Height		0.91 (0.74-0.98)	<0.001	0.88 (0.70-0.99)	<0.001		1.00 (0.99-1.01)	0.077	0.99 (0.98-1.01)	0.053
Weight		1.44 (1.13-1.51)	0.014	1.24 (1.01-1.42)	0.033		1.03 (1.03-1.04)	<0.001	1.02 (1.00-1.03)	<0.001
Smoking		1.77 (1.45-2.62)	0.017	1.52 (1.12-1.93)	0.028		1.76 (1.36-2.28)	<0.001	1.68 (1.33-2.12)	<0.001
SBP		1.13 (1.01-1.19)	0.024	1.03 (1.00-1.08)	0.015		2.37 (1.64-3.11)	<0.001	2.21 (1.55-2.93)	<0.001
DBP		1.17 (1.04-1.42)	<0.001	1.06 (1.01-1.15)	<0.001		2.01 (1.34-2.11)	<0.001	1.93 (1.36-2.21)	<0.001
eGFR		1.33 (1.10-1.51)	0.014	1.12 (1.03-1.39)	0.036		1.35 (1.21-1.45)	0.039	1.22 (1.12-1.43)	0.053
Blood glu.		3.32 (1.79-5.37)	<0.001	2.62 (1.48-4.62)	<0.001		4.06 (3.55-4.78)	<0.001	4.06 (3.55-4.78)	<0.001
AgeDiff		3.16 (2.11-5.28)	<0.001	2.74 (1.69-3.88)	<0.001		3.32 (2.37-4.14)	<0.001	2.45 (1.76-3.49)	<0.001
	CKD					Hypertension				
Sex		0.71 (0.53-0.93)	0.003	0.69 (0.64-0.72)	<0.001		0.93 (0.73-1.03)	<0.001	0.91 (0.75-1.02)	<0.001
BMI		1.04 (1.03-1.06)	<0.001	1.03 (1.02-1.06)	<0.001		1.21 (1.03-1.34)	<0.001	1.11 (1.04-1.21)	<0.001
Height		0.96 (0.93-0.99)	0.007	1.01 (1.00-1.03)	<0.001		0.74 (0.55-0.91)	0.077	0.73 (0.66-0.75)	0.033
Weight		1.06 (1.03-1.08)	0.014	1.00 (1.00-1.01)	0.033		1.26 (1.06-1.52)	<0.001	1.18 (1.13-1.31)	<0.001
Smoking		1.44 (1.15-1.61)	<0.001	1.32 (1.19-1.52)	0.028		1.74 (1.45-1.91)	<0.001	1.62 (1.39-1.72)	<0.001
SBP		1.55 (1.05-1.83)	<0.001	1.29 (1.02-1.43)	0.015		4.63 (2.45-6.48)	<0.001	4.31 (2.55-5.98)	<0.001
DBP		1.47 (1.13-1.62)	0.005	1.36 (1.15-1.55)	<0.001		3.28 (2.11-5.47)	<0.001	3.08 (2.33-4.84)	<0.001
eGFR		3.16 (2.60-3.51)	<0.001	3.37 (2.85-3.64)	0.036		1.21 (1.13-1.34)	0.039	1.22 (1.12-1.43)	0.062
Blood glu.		1.21 (0.99-1.31)	<0.001	1.07 (1.04-1.11)	<0.001		1.03 (1.00-1.05)	<0.001	1.02 (1.00-1.06)	0.031
AgeDiff		4.06 (3.55-4.78)	<0.001	4.14 (3.49-4.51)	<0.001		3.22 (2.46-3.76)	<0.001	3.11 (2.12-3.52)	<0.001
	CVD					Stroke				
Sex		0.82 (0.62-0.94)	0.005	0.71 (0.63-0.77)	0.011		1.02 (0.99-1.09)	0.005	0.71 (1.00-1.06)	0.011
BMI		1.24 (1.06-1.31)	0.004	1.14 (1.04-1.21)	0.014		1.02 (1.01-1.04)	0.003	1.03 (1.00-1.05)	0.012
Height		0.93 (0.88-0.97)	0.063	0.92 (0.91-0.99)	0.085		1.01 (1.00-1.03)	0.024	1.02 (1.00-1.04)	0.035
Weight		1.31 (1.05-1.48)	0.014	1.24 (1.01-1.42)	0.033		1.04 (1.00-1.08)	0.014	1.03 (1.01-1.06)	0.033
Smoking		1.84 (1.35-2.32)	<0.001	1.32 (1.19-1.52)	<0.001		1.54 (1.32-1.77)	<0.001	1.51 (1.30-1.64)	<0.001
SBP		1.55 (1.05-1.83)	0.024	1.29 (1.02-1.43)	0.035		1.45 (1.23-1.71)	0.014	1.29 (1.12-1.44)	0.023
DBP		1.47 (1.13-1.62)	0.017	1.36 (1.15-1.55)	0.043		1.11 (1.05-1.20)	0.014	1.36 (1.15-1.55)	0.043
eGFR		1.16 (1.01-1.31)	<0.001	1.09 (1.01-1.29)	<0.001		1.32 (1.21-1.44)	<0.001	1.15 (1.08-1.23)	<0.001
Blood glu.		2.52 (1.49-3.72)	<0.001	2.02 (1.68-3.01)	<0.001		2.13 (1.74-2.94)	<0.001	2.06 (1.74-2.64)	<0.001
AgeDiff		3.76 (2.25-4.93)	<0.001	3.14 (1.99-4.33)	<0.001		3.06 (2.25-4.93)	<0.001	3.14 (1.99-4.33)	<0.001

layers in MLP had 128 nodes and was applied with the rectified linear unit activation function. The Mean-Square Error loss was used as an objective function for the regression task of numerical value prediction between BA and CA. Other settings follow the default settings in Swin-Transformer V2 and BindFormer (45).

The multimodal fusion architecture training details were as follows. Transformations of random horizontal flip and rotations limited to ±20° were added to each batch during training as data augmentation to enable an improved and generalized network learning. We used AdamW optimizer (46) and cosine learning rate decay policy with an initial learning rate of 0.001. We used eight Telsa-A100 GPUs and trained the model for 500 epochs using Pytorch (47) library. The batch size was set to 256. We used five epochs for learning rate warm-up. We also used mixup and random augmentation techniques to boost the performance. We reported results on the test set using the optimal hyperparameters of our architecture selected in a grid search manner on the validation set.

3.4. Definition of AgeDiff and Criteria for Disease Diagnosis. We defined the AgeDiff between the predicted BA age using multimodal fusion method and CA, where a positive AgeDiff indicates a biological aging faster than the patient's CA, while a negative AgeDiff suggests that the BAs slower. The following criteria were used to define systemic diseases. CKD was defined as an estimated glomerular filtration rate (eGFR) of more than 60 mL min^{-1} per 1.73 m^2 with albuminuria or less than 60 mL min^{-1} per 1.73 m^2, confirmed in at least two visits separated by three months. Healthy controls were defined as eGFR above 60 mL min^{-1} per 1.73 m^2 without albuminuria, determined using a negative urine dip-stick test. Diabetes was defined by a fasting blood glucose $\geq$7.0 mmol L^{-1} at least two times, an HbA1c value of 6.5% or more and/or a history of drug treatment for diabetes. Hypertension was defined as a persistent increase in blood pressure above 130/80 or 140/90 mm Hg. Smoking as a risk factor was defined as participants smoke five cigarettes per day averagely.

3.5. Prediction of the Incidence Development of Systematic Diseases using Longitudinal Cohorts. For the incidence analysis of each disease, we denoted the index data as the time without disease (at baseline). The development of each disease was evaluated as an incidence data (or end point) within the yearly clinical follow-up. We trained the CPH models on the training and tuning set using variables based on the metadata and multimodal image–based risk score. The metadata-based model comprised sex, BMI, height, weight, smoking, systolic blood pressure (SBP), diastolic blood pressure (DBP), eGFR, and blood glucose. The multimodal image–based risk core is the predicted z-score (standard score) of the first visit generated from the detection model of each disease and used to predict progression risks of patients in combination with metadata. According to the risk scores of the first visit from the CPH model for the detection of each disease, the patients are triaged into three groups: low, medium, and high risk according to the upper and lower quartiles of predicted risk scores in the tuning set, respectively. Table 4 shows the distribution of the risk scores and the related thresholds (the upper and lower quartiles) across datasets. The risk scores were also treated as categorical variables according to quartiles during the incidence analysis on validation sets. Kaplan–Meier curves were constructed for the risk groups, and the significance of differences between group curves was computed using the log-rank test. Time-dependent ROC curves were used to quantify model performance on validation sets at the time of interest. ROC curves were constructed at a landmark time from predicted risk scores of relative patients made using the model. The univariable and multivariable CPH models were fitted. Two multivariable CPH models were developed, a combined metadata and fundus model and a metadata-only model serving as a baseline model. Statistical significance of HRs and adjusted HRs of CPH models were evaluated using the likelihood ratio test.

3.6. Contributions of Each Image Modality. To demonstrate the advantages of using three modalities on BA prediction and diagnostic predictions based on AgeDiff and study the contributions of each modality, we used the same implementation strategy with the same multimodal fusion framework in which we did suppression of each modality one by one, in order to examine the contributions of each of the three modalities. We fixed the architecture's parameters and used different training strategies to ensure a fair comparison until no further significant improvements were observed on the validation dataset. The main results on the test sets of two cohorts are shown in Table 5, which demonstrate that fundus images are the most important factor for an accurate prediction of BA (face and tongue weighted the second and the third in term of contribution) as well as analysis of related diseases, and a multimodal fusion of three modalities obtained the best performance. In addition, the results demonstrated that our multimodal fusion model is able to flexibly take an input with either one, two, or three modalities to produce an accurate BA estimation as well as a precise disease analysis.

3.7. Comparison with Other Deep Neural Networks. To demonstrate the advantages of our model, we also conducted experiments based on four popular deep learning models including two CNN-based ones (VGG-16 and ResNet-34) and two transformer-based ones (VIT-base-patch16-224 and Swin-tiny-patch4-window7-224), to compare with our multimodal fusion model. These models were pretrained on ImageNet and fine-tuned on the validation sets with late fusion strategy instead of CAM used in our implementations. Results were on the test sets of our two cohorts. From Table 6, we can observe that the BA prediction performance of simpler deep learning methods is around 5.5 MAE between BA and CA in the normal group and 7.0 in the abnormal group, while the numbers of our multimodal fusion architecture are more accurate with 2.0 MAE and 3.5 MAE, respectively. These results demonstrate superior performance of our multimodal fusion model on BA estimation and AgeDiff-based diagnostic predictions.

3.8. Interpretation of AI Predictions. We employed the Grad-CAM++ method (48) to produce visual explanations. Grad-CAM++ provides pixel-wise weighting of the gradients of the output with respect to a particular spatial position in any feature map of a DL-based system. In a single backward pass on the computational graph, a measure of importance of each pixel in a feature map toward the overall decision of the system was shown. In our scenario, the gradients of age difference between BA and CA were backpropagated through three MLP headers, MMT encoders, and linear projections to three input modalities. The saliency maps generated by Grad-CAM++ indicate the effect of each pixel on the model predictions. We applied Gaussian filtering to saliency maps for smoothness on three input modalities images. Fig. 3 shows an example of Grad-CAM++ results on three-modality inputs of one participant on internal training set in the training process. The saliency maps in the training process gradually provide visual clues on different regions of the face, fundus, and tongue (Table 7).

3.9. Statistical Analysis. To evaluate the performance of regression models for continuous values prediction (age) in this study, we calculated MAE, R^2 and PCC. We applied the Bland–Altman plot (49) to display the difference between CA

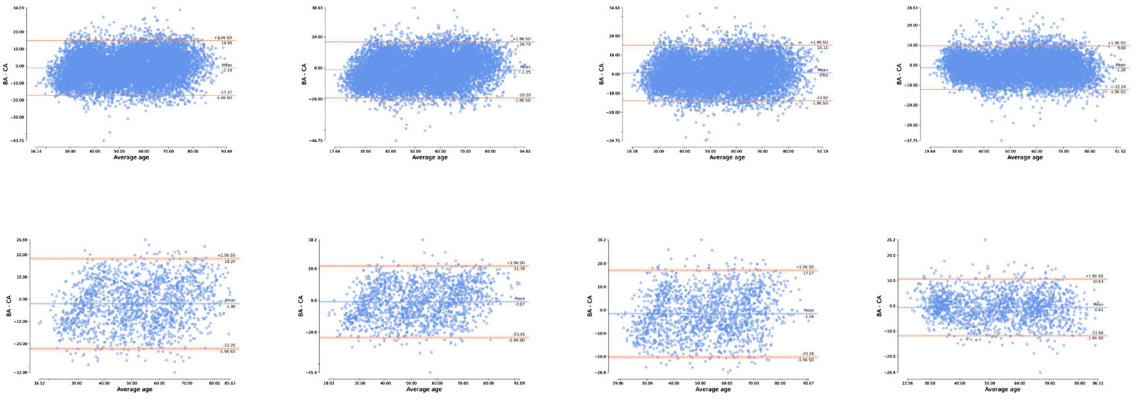

Fig. 8. Bland–Altman plots for the agreement between the predicted BA and CA on the internal test set and external test set (*Up*: face, tongue, fundus, and fusion from left to right). The *x* axis represents the mean of predicted BA and CA, and the *y* axis represents the difference between the two measurements.

and the predicted value of BA against the average of the two (Fig. 8). With 95% limits of agreement and ICC, we evaluated the agreement of the predicted BA and CA. We calculated the ratio between the variance of the model outputs and the variance of real-world data using the tuning set to calibrate outputs. Sensitivity and specificity were determined by the selected thresholds on the validation set. The models' performance on binary classification predictions was evaluated by ROC curves of sensitivity versus 1-specificity. The AUC of ROC curves was reported with 95% CI. The 95% CI of AUCs was estimated with the nonparametric bootstrap method (1,000 random resampling with replacement). The detection of each disease using BA was evaluated with binary classification models. We calculated the incidence rate for the whole cohort and for each risk group as the number of events per 1,000 person-years at risk. The Byar Poisson approximation method was used to calculate 95% CI of incidence (50). Then Kaplan–Meier estimators were constructed for different risk groups, and the significance of differences between groups was tested by log-rank tests. CPH models were tested using the likelihood ratio test. We used the time-dependent AUC at 4 y and 5 y to measure model performance. The Kaplan–Meier curve and the time-dependent ROC-AUC and PRC-AUC were calculated using the Python packages of lifelines (version 0.27.4) and scikit-survival (version 0.19.0).

Data, Materials, and Software Availability. The custom code is available at https://github.com/PKU-BDBA/BioAge (51). Restrictions apply to the availability of the developmental and validation datasets, which were used with permission of the participants for the current study. De-identified data may be available for research purposes from the corresponding authors on reasonable request.

ACKNOWLEDGMENTS. This research was supported by the Macau Science and Technology Development Fund, Macao (0007/2020/AFJ, 0070/2020/A2, and 0003/2021/AKP), MUST Faculty Research Grants (FRG-21-002-FMD), Guangzhou National Laboratory (YW-SLJC0201), Discipline Development of Peking University (7101302940 and 7101303005), Young Elite Scientist Sponsorship Program by Beijing Association for Science and Technology (BYESS2023026), National Key Research and Development Program of China (2021YFC2500500), and National Natural Science Foundation of China (6220071694, U22A20364, 81973112, and 81900626).

Author affiliations: [a]Department of Big Data and Biomedical AI, College of Future Technology, Peking University, Beijing 100871, China; [b]Macau Institute for AI in Medicine and Zhuhai People's Hospital and the First Affiliated Hospital of Faculty of Medicine, Macau University of Science and Technology, Macau 999087, China; [c]Guangzhou National Laboratory, Guangzhou 510005, China; [d]Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou 510623, China; [e]Dongguan People's Hospital, Southern Medical University, Dongguan 523059, China; [f]National Clinical Research Center for Ocular Diseases, Eye Hospital, Wenzhou Medical University, Wenzhou 325027, China; [g]Institute of Advanced Biotechnology and School of Life Sciences, Southern University of Science and Technology, Shenzhen 518055, China; [h]National Clinical Research Center for Kidney Diseases, State Key Laboratory for Organ Failure Research, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China; and [i]Clinical Research Institute, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 201620, China

1. L. Jia, W. Zhang, X. Chen, Common methods of biological age estimation. *Clin. Interv. Aging* **11**, 759–772 (2017).
2. J. Jylhävä, N. L. Pedersen, S. Hägg, Biological age predictors. *EBioMedicine* **21**, 29–36 (2017).
3. M. R. Hamczyk *et al.*, Biological versus chronological aging: JACC focus seminar. *J. Am. Coll. Cardiol.* **75**, 919–930 (2020).
4. A. Vaiserman, D. Krasnienkov, Telomere length as a marker of biological age: State-of-the-art, open issues, and future perspectives. *Front. Genet.* **11**, 630186 (2021).
5. G. Hannum *et al.*, Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
6. J. H. Cole *et al.*, Brain age predicts mortality. *Mol. Psychiatry* **23**, 1385–1392 (2018).
7. T. B. Brown *et al.*, Language models are few-shot learners. arXiv [Preprint] (2020). https://doi.org/10.48550/arXiv.2005.14165 (Accessed 28 May 2020).
8. Z. Zhu *et al.*, Retinal age gap as a predictive biomarker for mortality risk. *Br. J. Ophthalmol.* **107**, 547–554 (2023).
9. C. Liu *et al.*, "Biological age estimated from retinal imaging: A novel biomarker of aging" in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I* (Springer, 2019), pp. 138–146.
10. X. Xia *et al.*, Three-dimensional facial-image analysis to predict heterogeneity of the human ageing rate and the impact of lifestyle. *Nat. Metab.* **2**, 946–957 (2020).
11. W. Chen *et al.*, Three-dimensional human facial morphologies as robust aging markers. *Cell Res.* **25**, 574–587 (2015).
12. A. London, I. Benhar, M. Schwartz, The retina as a window to the brain–From eye research to CNS disorders. *Nat. Rev. Neurol.* **9**, 44–53 (2013).
13. C. Y. Cheung *et al.*, Deep-learning retinal vessel calibre measurements and risk of cognitive decline and dementia. *Brain Commun.* **4**, fcac212 (2022).
14. W. Hu *et al.*, Retinal age gap as a predictive biomarker of future risk of Parkinson's disease. *Age Ageing* **51**, afac062 (2022).
15. T. Imai, K. Okami, Facial cues to age perception using three-dimensional analysis. *PLoS One* **14**, e0209639 (2019).
16. Q. Liu *et al.*, A survey of artificial intelligence in tongue image for disease diagnosis and syndrome differentiation. *Digital Health* **9**, 20552076231191044 (2023).
17. T. Jiang *et al.*, Application of computer tongue image analysis technology in the diagnosis of NAFLD. *Comput. Biol. Med.* **135**, 104622 (2021).
18. Y. Li *et al.*, Oral, tongue-coating microbiota, and metabolic disorders: A novel area of interactive research. *Front. Cardiovasc. Med.* **8**, 730203 (2021).
19. C. Lu *et al.*, Oral-gut microbiome analysis in patients with metabolic-associated fatty liver disease having different tongue image feature. *Front. Cell. Infect. Microbiol.* **12**, 787143 (2022).
20. S. E. Kjeldsen, Hypertension and cardiovascular risk: General aspects. *Pharmacol. Res.* **129**, 95–99 (2018).
21. I. H. De Boer *et al.*, Diabetes and hypertension: A position statement by the American Diabetes Association. *Diabetes Care* **40**, 1273–1284 (2017).
22. G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree" in *Advances in Neural Information Processing Systems 30* (Curran Associates, Long Beach, California, USA, 2017).
23. S. M. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions" in *Advances in Neural Information Processing Systems 30* (Curran Associates, Long Beach, California, USA, 2017).
24. R. Chen *et al.*, Biomarkers of ageing: Current state-of-art, challenges, and opportunities. *MedComm Future Med.* **2**, e50 (2023), 10.1002/mef2.50.
25. S. Horvath, DNA methylation age of human tissues and cell types. *Genome biol.* **14**, 1–20 (2013).
26. J. H. Cole, K. Franke, Predicting age using neuroimaging: Innovative brain ageing biomarkers. *Trends Neurosci.* **40**, 681–690 (2017).
27. M. J. Peters *et al.*, The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* **6**, 1–14 (2015).
28. C. I. Weidner *et al.*, Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* **15**, 1–12 (2014).
29. J. G. Fleischer *et al.*, Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biol.* **19**, 1–8 (2018).
30. E. Putin *et al.*, Deep biomarkers of human aging: Application of deep neural networks to biomarker development. *Aging (Albany NY)* **8**, 1021 (2016).
31. P. Mamoshina *et al.*, Population specific biomarkers of human aging: A big data study using South Korean, Canadian, and Eastern European patient populations. *J. Gerontol. Series A* **73**, 1482–1490 (2018).
32. C.Y.-L. Cheung *et al.*, Imaging retina to study dementia and stroke. *Prog. Retin. Eye Res.* **57**, 89–107 (2017).
33. N. Patton *et al.*, Retinal vascular image analysis as a potential screening tool for cerebrovascular disease: A rationale based on homology between cerebral and retinal microvasculatures. *J. Anatomy* **206**, 319–348 (2005).
34. J. Cavanagh, H. Jones, Glycogenosomes in the aging rat brain: Their occurrence in the visual pathways. *Acta Neuropathol.* **99**, 496–502 (2000).
35. P.-C. Hsu *et al.*, Gender-and age-dependent tongue features in a community-based population. *Medicine* **98**, e18350 (2019).
36. R. B. Shaw Jr. *et al.*, Aging of the facial skeleton: Aesthetic implications and rejuvenation strategies. *Plast. Reconstr. Surg.* **127**, 374–383 (2011).
37. S. Kumar *et al.*, Teleophthalmology assessment of diabetic retinopathy fundus images: Smartphone versus standard office computer workstation. *Telemed. J. E. Health* **18**, 158–162 (2012).
38. K. Xia, J. Wang, Recent advances of transformers in medical image analysis: A comprehensive review. *MedComm Future Med.* **2**, e38 (2023), 10.1002/mef2.38.
39. Y. Wang *et al.*, China suboptimal health cohort study: Rationale, design and baseline characteristics. *J. Transl. Med.* **14**, 1–12 (2016).
40. N. Otsu, A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, pp.62–6266 (1979).
41. J. Ning *et al.*, Interactive image segmentation by maximal similarity based region merging. *Pattern Recognit.* **43**, 445–456 (2010).
42. J. A. Sethian, A fast marching level set method for monotonically advancing fronts. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1591–1595 (1996).
43. N. Sebkhi *et al.*, Evaluation of a wireless tongue tracking system on the identification of phoneme landmarks. *IEEE Transac. Biomed. Eng.* **68**, 1190–1197 (2020).
44. Z. Liu *et al.*, "Swin transformer v2: Scaling up capacity and resolution" in *IEEE Proceedings of the Conference on Computer Vision and Pattern Recognition* (New Orleans, Louisiana, USA, 2022), pp. 12009–12019.
45. G. Wang *et al.*, Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. *Nat. Med.* **29**, 2007–2018 (2023).
46. I. Loshchilov, F. Hutter, Decoupled weight decay regularization. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1711.05101 (Accessed 14 November 2017).
47. A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library" in *Advances in Neural Information Processing Systems 32* (Vancouver, Canada, 2019).
48. A. Chattopadhay *et al.*, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks" in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, Lake Tahoe, Nevada, USA, 2018), pp. 839–847.
49. D. Giavarina, Understanding bland altman analysis. *Biochem. Med.* **25**, 141–151 (2015).
50. N. E. Breslow, N. E. Day, *Statistical Methods in Cancer Research Volume II–the Design and Analysis of Cohort Studies* (IARC Scientific Publications, 1986), **vol. 1**.
51. J. Wang, K. Zhang, Multimodal-Biometric-Image. Github. https://github.com/PKU-BDBA/BioAge. Deposited 14 August 2023.