

Investigating the fundamental characteristics of retinal age models

Yukun Zhou

yukun.zhou.19@ucl.ac.uk

University College London <https://orcid.org/0000-0002-0840-6422>

Yiqun Lin

University College London <https://orcid.org/0000-0002-7697-0842>

Ariel Ong

Institute of Ophthalmology, University College London

Matthew Wong

University of Cambridge

Yilan Wu

University College London

Wenyi Hu

Centre for Eye Research Australia

Shubhank Verma

Institute of Ophthalmology, University College London

Broder Poschkamp

Institute for Community Medicine, University Medicine Greifswald

Lie Ju

Institute of Ophthalmology

Akshay Narayan

University College London

Sophie Martin

Department of Computer Science, University College London

Maitrei Kohli

Hawkes Institute, University College London

Fares Antaki

Cole Eye Institute, Cleveland Clinic

Alexander Heatley

University College London

Meng Wang

Yong Loo Lin School of Medicine, National University of Singapore

Ke Zou

National University of Singapore

Zheyuan Wang

Shanghai Jiao Tong University

Eden Ruffell

Institute of Ophthalmology, University College London

Dominic Williamson

Institute of Ophthalmology, University College London

Justin Engelmann

Institute of Ophthalmology, University College London, London, UK <https://orcid.org/0000-0002-5345-6023>

Hyunmin Kim

Institute of Ophthalmology, University College London

Takahiro Ninomiya

Graduate School of Medicine, Tohoku University

Rahul Jonas

Institute of Ophthalmology, University College London <https://orcid.org/0000-0002-5389-3047>

Yansong Liu

University College London

Zijie Cheng

University College London

Hanyuan Zhang

Department of Medical Physics and Biomedical Engineering, University College London

Hejie Cui

Stanford University School of Medicine

Hualiang Wang

The Hong Kong University of Science and Technology

Zuozhu Liu

Zhejiang University <https://orcid.org/0000-0002-7816-502X>

Bin Pu

Hunan University

Chubo Liu

Hunan University

Kenli Li

Hunan university

Jiayang Xu

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology

Xinpeng Ding

Hangzhou Institute of Technology, Xidian University

Guokai Zhang

University of Shanghai for Science and Technology

Huazhu Fu

Agency for Science, Technology and Research (A*STAR) <https://orcid.org/0000-0002-9702-5524>

Zongyuan Ge

Monash University <https://orcid.org/0000-0002-5880-8673>

Xujia Liu

Department of Ophthalmology and Visual Sciences, Faculty of Medicine, The Chinese University of Hong Kong

Mohammad Eslami

Department of Ophthalmology, Harvard Medical School

Milen Raytchev

Harvard Medical School

Tobias Elze

Department of Ophthalmology, Harvard Medical School

Michael Morley

Department of Ophthalmology, Harvard Medical School

Neil Oxtoby

University College London <https://orcid.org/0000-0003-0203-3909>

Daniel Alexander

Centre for Medical Image Computing, Department of Computer Science, University College London
<https://orcid.org/0000-0003-2439-350X>

Anthony Khawaja

Institute of Ophthalmology, University College London <https://orcid.org/0000-0001-6802-8585>

Ching-Yu Cheng

National University of Singapore <https://orcid.org/0000-0003-0655-885X>

Lisa Zhu

Centre for Eye Research Australia; Ophthalmology, Department of Surgery, University of Melbourne
<https://orcid.org/0000-0002-9897-1192>

Yih-Chung Tham

National University of Singapore <https://orcid.org/0000-0002-6752-797X>

James Cole

UCL <https://orcid.org/0000-0003-1908-5588>

Carol Cheung

The Chinese University of Hong Kong <https://orcid.org/0000-0002-9672-1819>

Pearse Santos

NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust
<https://orcid.org/0000-0002-9239-745X>

Siegfried Wagner

University College London Institute of Ophthalmology, University College London
<https://orcid.org/0000-0003-4915-4353>

Article

Keywords:

Posted Date: July 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10157626/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

1 **Investigating the fundamental characteristics of retinal age models**

2 **Authors:** Yiqun Lin^{1,2,3,4}, Ariel Yuhan Ong^{1,2}, Matthew Yu Heng Wong⁵, Yilan Wu^{1,6}, Wenyi Hu^{7,8},
3 Shubhank Shobhit Verma¹, Broder Poschkamp^{1,9}, Lie Ju^{1,2}, Akshay Narayan¹⁰, Sophie A. Martin^{3,4},
4 Maitrei Kohli^{3,4}, Fares Antaki^{11,12,13}, Alexander C. Heatley¹, Meng Wang^{14,15}, Ke Zou^{14,16}, Zheyuan
5 Wang^{17,18}, Eden Ruffell^{1,2}, Dominic Williamson^{1,2}, Justin Engelmann¹, Hyunmin Kim^{1,2}, Takahiro
6 Ninomiya^{1,19}, Rahul A. Jonas¹, Yansong Liu⁴, Zijie Cheng¹⁰, Hanyuan Zhang^{3,20}, Hejie Cui²¹, Hualiang
7 Wang²², Zuozhu Liu^{23,24}, Bin Pu²⁵, Chubo Liu²⁵, Kenli Li²⁵, Jiayang Xu²², Xinpeng Ding²⁶, Guokai
8 Zhang²⁷, Huazhu Fu²⁸, Zongyuan Ge^{29,30}, Xujia Liu³¹, Mohammad Eslami^{32,33}, Milen Raytchev^{32,33},
9 Tobias Elze^{32,33}, Michael G. Morley^{32,33}, Neil P. Oxtoby^{3,4}, Daniel C. Alexander^{3,4}, Anthony P. Khawaja^{1,2},
10 Ching-Yu Cheng^{14,15}, Lisa Zhuoting Zhu^{7,8}, Yih Chung Tham^{14,15,16}, James H. Cole^{3,4}, Carol Y. Cheung³¹,
11 Pearse A. Keane^{1,2}, Siegfried K. Wagner^{1,2}, Yukun Zhou^{1,2,3,4+}

12

13 **Affiliations:**

14 ¹ Institute of Ophthalmology, University College London, London, UK

15 ² NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK

16 ³ UCL Hawkes Institute, University College London, London, UK

17 ⁴ Department of Computer Science, University College London, London, UK

18 ⁵ University of Cambridge, Cambridge, UK

19 ⁶ Beijing Visual Science and Translational Eye Research Institute, Eye Center of Beijing Tsinghua
20 Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing,
21 China

22 ⁷ Centre for Eye Research Australia, Melbourne, VIC, Australia

23 ⁸ Department of Surgery, Melbourne Medical School, The University of Melbourne, Melbourne, VIC,
24 Australia

25 ⁹ Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany

26 ¹⁰ University College London, London, UK

27 ¹¹ Cole Eye Institute, Cleveland Clinic, Cleveland, OH, USA

28 ¹² Department of Ophthalmology, Centre Hospitalier de l'Université de Montréal, Montréal, QC, Canada

29 ¹³ The CHUM School of Artificial Intelligence in Healthcare, Centre Hospitalier de l'Université de
30 Montréal, Montreal, QC, Canada

31 ¹⁴ Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore,
32 Singapore

33 ¹⁵ Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore

34 ¹⁶ Centre for Innovation and Precision Eye Health, National University of Singapore, Singapore

35 ¹⁷ Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

36 ¹⁸ MOE Key Laboratory of Artificial Intelligence, School of Electronic, Information, and Electrical
37 Engineering, Shanghai Jiao Tong University, Shanghai, China

38 ¹⁹ Graduate School of Medicine, Tohoku University, Sendai, Japan

39 ²⁰ Department of Medical Physics and Biomedical Engineering, University College London, London, UK

- 40 ²¹ Stanford University School of Medicine, Stanford, CA, USA
- 41 ²² Department of Electronic and Computer Engineering, The Hong Kong University of Science and
42 Technology, Hong Kong SAR, China
- 43 ²³ ZJU-Angelalign R&D Center for Intelligent Healthcare, ZJU-UIUC Institute, Zhejiang University,
44 Haining, China
- 45 ²⁴ Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Zhejiang University, Hangzhou,
46 China
- 47 ²⁵ College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
- 48 ²⁶ Hangzhou Institute of Technology, Xidian University, Hangzhou, China
- 49 ²⁷ University of Shanghai for Science and Technology, Shanghai, China
- 50 ²⁸ Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR),
51 Singapore
- 52 ²⁹ Department of Data Science and AI, Monash University, Melbourne, VIC, Australia
- 53 ³⁰ AIM for Health Lab, Monash University, Melbourne, VIC, Australia
- 54 ³¹ Department of Ophthalmology and Visual Sciences, Faculty of Medicine, The Chinese University of
55 Hong Kong, Hong Kong SAR, China
- 56 ³² Department of Ophthalmology, Harvard Medical School, Boston, MA, USA
- 57 ³³ Schepens Eye Research Institute of Massachusetts Eye and Ear, Boston, MA, USA
- 58
- 59 ⁺ **Corresponding author:** Yukun Zhou (yukun.zhou.19@ucl.ac.uk)
- 60

61 **Abstract**

62 Retinal age is a biological ageing marker estimated from retinal images. Its deviation from chronological
63 age, termed the retinal age gap (RAG), has been associated with adverse health outcomes. However,
64 fundamental characteristics of retinal age models that are critical for interpreting and applying RAG
65 remain underexplored, including the consistency of these associations across age subgroups as well as
66 model generalisability across different cohorts. In this study, we used 327,764 retinal images to develop
67 a retinal age model and investigate associations between RAG and systemic diseases. We examined
68 association consistency across age subgroups and evaluated model performance on three external
69 datasets with distinct imaging devices and populations. RAG showed positive associations with systemic
70 diseases in young and middle-aged subgroups, but negative associations in the aged subgroup,
71 revealing substantial age-dependent inconsistency. We further showed that this inconsistency was
72 explained by regression-to-the-mean effects which varied by health status. In external evaluations, the
73 generalisability of RAG varied considerably by cohort and intended application. Large age estimation
74 errors did not necessarily indicate limited clinical utility, instead suggesting that model generalisability
75 should be defined and evaluated in an application-specific manner rather than based on age estimation
76 accuracy alone. Overall, this study investigates key characteristics of retinal age models that affect the
77 interpretation of RAG and its disease associations. Our findings highlight the importance of reporting
78 age-dependent associations, and emphasise the need for multi-dimensional, application-specific
79 external evaluation for reliable use of retinal age and RAG in future clinical applications. More broadly,
80 our findings and practical recommendations may extend to wider biological age estimation models.

81

82 Introduction

83 Ageing is a complex biological process characterised by the accumulation of cellular damage and
84 physiological decline [1,2]. While chronological age reflects the time since birth, it does not fully capture
85 an individual's true health status, which is influenced by a combination of factors such as socioeconomic
86 determinants, genetic background, and comorbidity burden [3,4]. In contrast, biological ageing markers,
87 such as brain age [5–9], face age [10], and epigenetic clocks [11,12] better reflect physiological status
88 predictive of health outcomes [7,13–16]. However, the measurement of many of these established
89 biomarkers is constrained by high cost, invasiveness, or complex resource requirements [17,18].
90 Recently, retinal age has emerged as a particularly promising and non-invasive alternative [15,19–25]
91 that can be directly estimated from retinal images.

92
93 The retina offers a unique window into the body, as it shares embryological origins and physiological
94 properties with the central nervous and vascular systems. It allows for the direct visualisation of blood
95 vessels and neural tissue [15,26,27]. Recent research confirms that retinal images contain crucial
96 information about overall health, effectively identifying or predicting systemic conditions such as
97 cardiovascular disease and neurodegenerative disease [28–31]. Because retinal imaging is rapid and
98 non-invasive [28], it is increasingly used to quantify biological ageing [25,29]. To establish a reference
99 standard for biological age in model development, researchers have explored the utility of complex
100 molecular targets such as epigenetic clocks [11,12], transcriptomic signatures [32], and proteomic
101 profiles [16,33,34]. While these high-dimensional molecular markers offer mechanistic insights into
102 cellular ageing, their acquisition is often costly and requires specialised biological sampling [24].
103 Consequently, using chronological age as the primary target label [7,15,17] remains the most accessible
104 approach, given its wide availability in large cohorts.

105
106 As illustrated in [Fig. 1a](#), retinal age models utilise retinal images as inputs and output an estimated
107 retinal age. Previous studies assume that biological age approximates chronological age in a relatively
108 healthy population [17,19,35,36]. Under this assumption, an estimated retinal age older than
109 chronological age likely indicates poorer health. In clinical association studies, the discrepancy between
110 estimated retinal age and chronological age, defined as the retinal age gap (RAG), serves as a direct
111 indicator of general health status [15,20]. RAG has shown significant associations with all-cause
112 mortality [15], cardiovascular events [20,23], Parkinson's disease [22], kidney failure [37], schizophrenia
113 [38], and multimorbidity [39], illustrating its potential as a screening tool [17,36]. Previous studies have
114 utilised artificial intelligence (AI) techniques to estimate biological age from retinal images on healthy
115 cohorts [17,19,35,36]. While early approaches primarily relied on standard convolutional neural
116 networks to extract localised anatomical features [15,19], subsequent research introduced Vision
117 Transformers (ViTs) to better capture complex, global contextual representations across the retina [40].
118 More recently, retinal foundation models like RETFound [41] have been widely used due to their strong
119 generalisability and representational capacity across retinal imaging tasks. They are increasingly being
120 used for precise retinal age estimation [42,43].

121
122 However, retinal age models are trained to predict chronological age, which can lead their predictions to
123 be systematically pulled towards the mean age of the training population, a phenomenon known as
124 regression-to-the-mean (RTM) in regression models [44–47]. The RTM effect can cause age to be
125 overestimated in young individuals and underestimated in old individuals, with estimated ages being

126 more concentrated around the mean age, as illustrated in Fig. 1b. For retinal age estimation, RTM is
127 particularly relevant because RAG is widely used in association studies to interpret health status
128 [15,17,20,21,25], as illustrated in Fig. 1a. Because RTM systematically shifts estimated ages towards
129 the population mean, the resulting RAG may partially reflect age-related statistical bias, potentially
130 leading to inconsistent associations across age subgroups. Despite RAG's potential impact on clinical
131 interpretation, this issue and its associated clinical implications have not been fully examined in previous
132 studies.

133
134 In this study, we investigated fundamental characteristics of retinal age models and examined their
135 implications for retinal age interpretation and clinical applications. We developed a retinal age model
136 and conducted association analyses using the AlzEye dataset [48] (Fig. 1c). Consistent with previous
137 studies [15,17,20,21,25], our results showed that a higher RAG is positively associated with poorer
138 health status in the overall cohort. However, we identified differential RTM effects between healthy and
139 unhealthy groups (Fig. 1b) and observed substantially varied associations between RAG and health
140 conditions across age subgroups. Specifically, the positive association was attenuated or even reversed
141 to negative in the aged subgroup. We further demonstrated that incorporating these age-dependent
142 associations into the modelling process improved disease classification. Additionally, we evaluated the
143 retinal age model on three external datasets - the UK Biobank, BRSET, and mBRSET (Fig. 1c). External
144 evaluation showed that the generalisability of RAG varied largely by cohort and intended application, as
145 assessed across multiple dimensions, including age estimation accuracy, disease association strength,
146 and disease detection performance. Our findings highlight the importance of reporting age-dependent
147 associations. Building on our findings, we propose practical strategies for using RAG in disease
148 detection, and emphasise the need for multi-dimensional, application-specific external evaluation to
149 support more reliable use in future clinical research.

150

151 **Results**

152 **Study cohort**

153 RAG was defined as estimated retinal age minus chronological age [15,17,22,23]. We used the AlzEye
154 [48] dataset for retinal age model development and internal evaluation, and three additional cohorts for
155 external evaluation: UK Biobank [49], BRSET [50] and mBRSET [51]. The internal AlzEye cohort
156 comprised 60,865 patients and 327,764 macula-centred colour fundus photographs (CFPs). In this
157 study, the definitions of healthy and unhealthy status were relative and depended on the disease
158 records available in the curated cohort. Participants were classified as unhealthy if they had one or more
159 confirmed disease records, including cardiovascular diseases (atrial fibrillation, heart failure, myocardial
160 infarction, intracerebral haemorrhage, ischaemic stroke, aortic stenosis, aortic regurgitation, mitral
161 stenosis, and mitral regurgitation), chronic systemic diseases (diabetes mellitus and hypertension),
162 ocular diseases (glaucoma, diabetic retinopathy, and age-related macular degeneration), and
163 neurodegenerative diseases (dementia and Parkinson's disease); all remaining participants were
164 classified as healthy in this study. The curated AlzEye cohort had a mean age of 63.07 ± 11.76 years.
165 Data curation details are described in the Methods section. Internal cohort curation, dataset splits, and
166 age distributions are summarised in [Supplementary Fig. 1-3](#).

167
168 External evaluation included three independent cohorts with distinct study designs, populations and
169 imaging devices. The UK Biobank, a large prospective epidemiological study from the United Kingdom,
170 contributed 29,835 patients and 51,709 images after curation, all acquired using Topcon devices, with a
171 mean age of 56.05 ± 8.23 years. BRSET and mBRSET were both Brazilian cohorts but differed in
172 imaging acquisition: BRSET included 5,529 patients and 10,287 images acquired with desktop Canon
173 CR-2 and Nikon NF505 cameras, with a mean age of 57.09 ± 18.20 years, whereas mBRSET included
174 1,277 patients and 5,107 images acquired with the handheld Phelcom Eyer camera, with a mean age of
175 61.41 ± 11.41 years. Data curation details are described in the Methods section. External cohort curation,
176 dataset splits, and age distributions are summarised in [Supplementary Fig. 4-8](#).

177 **Retinal age and disease associations on the whole internal set**

178 The retinal age model was trained on the AlzEye cohort ([Fig. 2a](#)) and achieved a mean absolute error
179 (MAE) of 5.31 years, a root mean square error (RMSE) of 6.80 years, and a squared Pearson
180 correlation (Pearson's r^2) of 0.745 between estimated and chronological age in internal evaluation ([Fig.](#)
181 [2b](#)). In [Fig. 2c](#), logistic regression analyses showed positive associations between RAG and disease
182 status after adjustment for chronological age. Each 1-year increase in RAG was associated with higher
183 odds of being in the unhealthy group (odds ratio (OR) = 1.049, 95% CI: [1.046, 1.053]), cardiovascular
184 disease group (OR = 1.041, 95% CI: [1.036, 1.046]), chronic systemic disease group (OR = 1.048, 95%
185 CI: [1.045, 1.052]) and ocular disease group (OR = 1.065, 95% CI: [1.061, 1.069]). All associations were
186 significant (all $p < 0.001$). These results show that the retinal age model achieved reasonable internal
187 estimation performance and reproduced the conventional positive association between higher RAG and
188 diseases.

189 **Inconsistent association caused by differential RTM**

190 As shown in Fig. 1b, the associations between RAG and disease varied across age subgroups. Positive
191 associations were observed in young and middle-aged cohorts ($OR > 1$), whereas a negative association
192 was observed in the older cohort ($OR < 1$). We explored the underlying cause of this age-dependent
193 inconsistency. In Fig. 3a, we visualised the relationship between RAG and chronological age across
194 healthy and unhealthy groups. RAG declined with age in both groups, but the unhealthy group showed a
195 visibly steeper slope. In Fig. 3b, we fit an ordinary least squares linear model of RAG on chronological
196 age, health status, and their interaction to estimate the RAG-age slope. The healthy slope was -0.184
197 (95% CI: [-0.191, -0.177]) and the unhealthy slope was -0.314 (95% CI: [-0.318, -0.310]), with an
198 interaction coefficient $p < 0.001$. In Fig. 3c and Supplementary Fig. 10, we extended this analysis to
199 disease categories and found that cardiovascular and ocular diseases showed steeper slopes of -0.345
200 (95% CI: [-0.353, -0.336]) and -0.345 (95% CI: [-0.352, -0.339]). In addition, participants with more
201 disease categories showed progressively steeper RAG-age slopes, from -0.289 in those with one
202 disease category to -0.356 and -0.390 in those with two and three disease categories, respectively
203 (Supplementary Fig. 11). This steeper slope in disease groups relative to the healthy group reflected
204 differential RTM, leading to inconsistent associations between RAG and unhealthy conditions across
205 age subgroups (Fig. 1b). This contrasts with previous studies that often treated the association as
206 constant over the whole cohort.

207
208 We explored whether differential RTM could be removed by calibrating RAG to chronological age. In
209 Fig. 3d,e, we applied linear and quadratic age-calibrators fitted on the healthy subset of the downstream
210 training set. Both calibrations largely flattened the RAG-age trend in the healthy group: the healthy slope
211 was 0.017 (95% CI: [0.011, 0.024]) after linear calibration and 0.018 (95% CI: [0.011, 0.024]) after
212 quadratic calibration. However, a significant negative slope persisted in the unhealthy group under both
213 approaches: -0.113 (95% CI: [-0.117, -0.109]) after linear calibration and -0.101 (95% CI: [-0.105, -
214 0.097]) after quadratic calibration. Regardless of the calibration method used, the slope difference
215 between the two groups (quantified by the Age $\times$ Group interaction term) remained significant
216 ($p < 0.001$), still leading to inconsistent associations. These results indicate that age calibration can
217 remove the average RAG-age trend in the healthy group, but cannot remove differential RTM between
218 healthy and unhealthy groups.

219
220 We tested whether differential RTM could be mitigated by balancing the age distribution of the training
221 data. In Fig. 3f, we constructed a training set by evenly sampling over ages to train a retinal age model.
222 In Fig. 3g, we evaluated the trained model on the internal test set. The differential RTM still existed: the
223 healthy slope was -0.190 (95% CI: [-0.196, -0.183]) and the unhealthy slope was -0.344 (95% CI: [-
224 0.348, -0.340]), with the slope difference between the two groups remained significant ($p < 0.001$). Thus,
225 balanced sampled data cannot address the differential RTM across groups, causing the associations
226 between RAG and unhealthy conditions to still vary with age.

227 **Added value of RAG for disease classification**

228 Acknowledging the different RTM and inconsistent associations across age subgroups, we incorporated
229 an interaction term between age and RAG (Age $\times$ RAG) to capture the dynamics in the modelling. In
230 Fig. 4a, the Age + RAG + Age $\times$ RAG model showed that the odds ratio for unhealthy status per 1-year

231 increase in RAG declined with chronological age. The odds ratio remained above 1 in the young age
232 subgroup and the middle age subgroup, indicating that higher RAG was associated with higher odds of
233 unhealthy status. By contrast, in the old age subgroup (>90 years), the odds ratio approached or fell
234 below 1, indicating that the disease association of RAG was substantially weakened or reversed at very
235 old ages.

236
237 In Fig. 4b, we compared three logistic regression models for disease classification: Age only, Age +
238 RAG, and Age + RAG + Age $\times$ RAG. Across all four disease outcomes, the interaction model achieved
239 the highest area under the receiver operating characteristic curve (AUC), with AUCs of 0.741 (95% CI:
240 [0.736, 0.746]) for unhealthy conditions, 0.813 (95% CI: [0.808, 0.818]) for cardiovascular disease,
241 0.737 (95% CI: [0.732, 0.742]) for chronic disease, and 0.774 (95% CI: [0.769, 0.780]) for ocular
242 disease. These corresponded to AUC improvements of 0.7-1.6% over the Age-only model and
243 additional improvements of 0.4-0.5% over the Age + RAG model across the four outcomes. Although the
244 performance improvement is modest, these indicate the added value of RAG and the interaction term
245 (Age $\times$ RAG).

246
247 In Fig. 4c, we evaluated model performance within age subgroups. For the unhealthy outcome, the Age-
248 only model achieved AUCs of 0.580 (95% CI: [0.567, 0.594]) in the young age subgroup (<50 years),
249 0.622 (95% CI: [0.616, 0.629]) in the middle age subgroup (50-80 years), and 0.585 (95% CI: [0.566,
250 0.603]) in the older age subgroup (>80 years). Adding RAG improved performance in the young and
251 middle age subgroups, increasing AUC to 0.674 (95% CI: [0.662, 0.686]) and 0.641 (95% CI: [0.634,
252 0.648]), respectively. However, in the older age subgroup, the Age + RAG model performed worse than
253 the Age-only model, with AUC decreasing to 0.494 (95% CI: [0.475, 0.512]). Further including the Age $\times$
254 RAG interaction term maintained the performance gains in the young and middle age subgroups, with
255 AUCs of 0.685 (95% CI: [0.672, 0.697]) and 0.641 (95% CI: [0.635, 0.649]), while substantially reducing
256 the performance degradation in the older age subgroup, achieving an AUC of 0.564 (95% CI: [0.546,
257 0.582]). In Fig. 4d, adding RAG and the Age $\times$ RAG interaction progressively refined the decision
258 boundary, allowing previous misclassifications to be corrected by the new threshold.

259 **Age-estimation error, disease association strength, and value for disease classification on**
260 **external evaluation**

261 We evaluated external performance across UK Biobank, BRSET and mBRSET. The AlzEye-trained
262 retinal age model was applied to each external test set without retraining or fine-tuning. Retinal age
263 estimation performance varied substantially across cohorts (Fig. 5a). Pearson's r^2 was 0.716 in UK
264 Biobank, 0.783 in BRSET and 0.432 in mBRSET, with corresponding MAE of 3.80, 7.34 and 8.40 years
265 and RMSE of 4.87, 9.37 and 10.52 years, respectively. These results indicate that the model retained
266 reasonable age-estimation performance in the UK Biobank but showed larger estimation errors in
267 BRSET and mBRSET.

268
269 We examined the associations between RAG and unhealthy conditions in external cohorts, as shown in
270 Fig. 5b and Supplementary Tab. 11. In the UK Biobank, the odds ratio was stable across age, ranging
271 from 1.026 at age 40 to 1.028 at age 70, with all available estimates significant ($p < 0.001$). In BRSET,
272 the odds ratio gradually decreased with age but remained significantly positive from age 30 to 90,

ranging from 1.052 at age 30 to 1.024 at age 90 ($p < 0.001$). In mBRSET, the odds ratio increased with age, from 1.008 at age 30 to 1.058 at age 80; the association was not significant at age 30 or 40 ($p > 0.05$) but became significant from age 50 onward ($p < 0.001$). These findings suggest that, despite cohort differences in age estimation error, the positive associations between RAG and unhealthy conditions generally remained significant in the external cohorts.

We assessed external disease classification under off-the-shelf and externally calibrated settings. As shown in [Fig. 5c](#), in AlzEye, the interaction model (Age + RAG + Age $\times$ RAG) achieved a higher AUC than the Age-only model, increasing from 0.731 (95% CI: [0.706, 0.756]) to 0.741 (95% CI: [0.715, 0.767]; $p < 0.001$ for AUC difference). In UKB, the AUC difference was not significant under off-the-shelf evaluation, with AUCs of 0.684 (95% CI: [0.674, 0.693]) for the Age-only model and 0.685 (95% CI: [0.675, 0.694], $p = 0.812$) for the interaction model. After external calibration, the interaction model increased AUC from 0.684 (95% CI: [0.674, 0.693]) to 0.688 (95% CI: [0.679, 0.697]; $p < 0.001$). In BRSET, the AUC difference was not significant under off-the-shelf evaluation, with AUCs of 0.662 (95% CI: [0.644, 0.680]) for the Age-only model and 0.664 (95% CI: [0.646, 0.682], $p = 0.508$) for the interaction model. After external calibration, the interaction model increased AUC from 0.662 (95% CI: [0.644, 0.680]) to 0.671 (95% CI: [0.652, 0.689]; $p < 0.05$). In mBRSET, the AUC difference was not significant under off-the-shelf evaluation, with AUCs of 0.638 (95% CI: [0.582, 0.695]) for the Age-only model and 0.614 (95% CI: [0.559, 0.665], $p = 0.150$) for the interaction model. After external calibration, the interaction model showed similar AUC to the Age-only model, with AUCs of 0.638 (95% CI: [0.582, 0.695]) and 0.638 (95% CI: [0.583, 0.695], $p = 0.938$), respectively. Threshold-based metrics ([Fig. 5d](#)) also differed across cohorts, likely reflecting differences in healthy-unhealthy class distribution and the cohort-specific operating thresholds ([Supplementary Fig. 2, 6-9](#)). Overall, these results indicate that the added value of RAG for disease classification differed across external cohorts, suggesting that age-estimation accuracy and disease association strength alone are insufficient to determine its utility in downstream disease classification.

300 **Discussion**

301 In this study, we investigated key characteristics of retinal age models that are essential for their clinical
302 interpretation and applications, including the consistency of association between RAG and health
303 conditions across age subgroups, as well as the generalisability of retinal age models across cohorts
304 and applications. Our results confirm that RAG contains health-related information, while also revealing
305 age-dependent associations biased by differential RTM effects between healthy and unhealthy cohorts.
306 We further explored the benefits of incorporating these age-dependent associations into the modelling
307 process for disease detection. Additionally, our external evaluation showed that the generalisability of
308 RAG varied largely by cohort and intended application, indicating that RAG generalisability should be
309 assessed according to its intended use, rather than by a single evaluation metric.

310
311 Our retinal age model reproduced the expected positive association between higher RAG and poorer
312 health conditions in the overall cohort. The model achieved relatively low error in retinal age estimation.
313 After adjustment for chronological age, each 1-year increase in RAG was associated with higher odds of
314 being unhealthy. Similar positive associations were observed for cardiovascular disease, chronic
315 systemic disease and ocular disease (Fig. 2c). Our observations are consistent with previous studies
316 [15–17,20,21,25] identifying higher RAG to adverse health outcomes. These findings suggest that the
317 model does not merely estimate chronological age but also captures underlying pathological features
318 and morphological manifestations related to health status, such as vascular alterations [52,53].

319
320 Our analysis revealed inconsistent associations between RAG and unhealthy conditions across age
321 subgroups, caused by differential RTM across healthy and unhealthy groups. In younger age
322 subgroups, higher RAG was positively associated with unhealthy conditions, while this positive
323 association attenuated with increasing chronological age and could even become negative in older age
324 subgroups (Fig. 4a-b). This age-dependent inconsistency is likely driven by differential RTM between
325 the healthy and unhealthy groups. Disease-related lesions and pathological manifestations in the
326 unhealthy cohort likely introduce more challenging cases for retinal age models trained predominantly
327 on relatively healthy data, pulling predictions more strongly towards the population mean age,
328 determined by the training objective of regression models. Differential RTM relative to the healthy group
329 was more substantial in the cardiovascular and ocular groups than in the unhealthy and chronic groups.
330 This may be because ocular diseases can directly introduce visible pathological changes in retinal
331 images, while cardiovascular diseases involved in our study reflect more severe systemic conditions and
332 therefore clearer retinal changes. Consistently, groups with more disease categories showed more
333 pronounced differential RTM relative to the healthy group (Supplementary Fig. 11), further suggesting
334 that RTM may become more severe when retinal images contain pathological changes from multiple
335 disease categories.

336
337 The above findings warrant the incorporation of age-dependent associations into the modelling process
338 for more accurate interpretation and disease detection. These inconsistent associations suggest that
339 reporting a single overall RAG-disease association, as commonly done in previous studies
340 [15,17,20,21,25,54], may not fully capture the relationship between RAG and health status across
341 different age groups. We further found that differential RTM could not be removed by applying the same
342 age calibration of RAG to both groups (Fig. 3d,e). Additionally, calibration performed separately on the
343 healthy and unhealthy groups would remove the health-related signals that RAG is expected to capture.

344 We also found that balancing the training-age distribution did not resolve differential RTM. These
345 findings suggest that differential RTM is not a simple calibration artefact or a consequence of the
346 training-age distribution alone, but a fundamental characteristic of retinal age models that should be
347 explicitly considered when interpreting and applying RAG in clinical studies. In many RAG-based
348 association studies and disease classification [15,17,21,23], RAG is commonly used as an independent
349 predictor, without explicitly considering its interaction with age. However, our results suggest that this
350 may decrease the performance of disease classification in older age subgroups, relevant to the reversed
351 association between RAG and unhealthy conditions. Incorporating the Age $\times$ RAG interaction term
352 allowed the model to account for this age-dependent association, further improving the overall added
353 value of RAG (Fig. 4b) for disease classification. More importantly, in older age subgroups, the
354 interaction term prevented the value of RAG from becoming negative and mitigated the misleading risk
355 interpretation (Fig. 4c) potentially caused by adding RAG alone, which assumes that higher RAG
356 uniformly indicates poorer health across age groups. Although the AUC gain of RAG in disease
357 classification was modest, these results support the presence of additional information from RAG. These
358 findings indicate that RAG remains informative despite differential RTM, but clinical modelling should
359 account for age-dependent associations in downstream applications. More broadly, it is worthwhile to
360 investigate association consistency linking biological ageing markers with systemic conditions, before
361 reporting a single value association across the overall cohort.

362
363 Our external evaluations suggest that the generalisability of retinal age models should be assessed in
364 the context of their intended applications. Importantly, performance metrics such as age-estimation
365 accuracy, disease association strength, and value for disease classification may not improve or
366 deteriorate in parallel, highlighting the need for application-specific evaluation frameworks. While
367 previous studies [17,24,35,40] have primarily focused on external age-estimation performance and
368 disease associations, our results suggest that these metrics are informative for different applications and
369 should be interpreted accordingly. For applications that require accurate age estimation, external age-
370 estimation performance remains a key evaluation criterion. In the UK Biobank, the model maintained
371 relatively good age-estimation performance, indicating satisfactory generalisability for this purpose.
372 However, poor age-estimation accuracy did not necessarily preclude utility in other settings. In both
373 BRSET and mBRSET, substantial age-estimation errors were observed, yet RAG remained significantly
374 associated with unhealthy conditions across much of the age range (Supplementary Tab. 11). One
375 possible explanation is that domain shifts introduced a cohort-level offset in retinal age estimates,
376 reducing absolute age-estimation accuracy while preserving disease-related variation in RAG. For
377 applications that use retinal age as a biomarker of health status, preserving disease associations may
378 be more important than achieving highly accurate age estimates. Nevertheless, significant associations
379 alone were insufficient to establish clinical utility in downstream predictive tasks. In BRSET, RAG
380 provided additional information beyond chronological age for disease classification despite the
381 substantial estimation error. In contrast, in mBRSET, the improvement in classification performance
382 beyond chronological age was small without statistical significance, despite significant disease
383 associations (Supplementary Tab. 12). This suggests that biomarkers showing population-level
384 associations may not necessarily provide sufficient discriminative information for disease prediction.
385 Taken together, these findings indicate that the generalisability of retinal age models should be
386 evaluated against the requirements of their intended application. Depending on whether the goal is age
387 estimation, association analysis, risk stratification, or disease classification, different evaluation criteria

388 may be most relevant, and strong performance in one application should not be assumed to translate
389 directly to another.

390
391 The RTM [44–46] and limited generalisability observed here are unlikely to be unique to retinal age
392 models, but may represent broader challenges in biological age estimation. Similar age-dependent
393 biases have been reported in MRI-derived brain age models, where younger individuals tend to be
394 overestimated and older individuals underestimated, resulting in systemic associations between brain
395 age gap and chronological age [45–47,55]. Epigenetic clocks exhibit comparable biases, including
396 systematic underestimation in older individuals or tissues, which can influence the interpretation of age-
397 acceleration measures [56]. Likewise, limited generalisability across populations and acquisition settings
398 has been reported in both brain age models and epigenetic clocks, arising from factors such as scanner
399 and imaging-protocol differences in MRI [57] and batch or assay-platform effects in DNA methylation
400 data [16,58]. Our findings extend these observations by showing that RTM effects can differ between
401 healthy and unhealthy groups, leading to age-dependent associations between biological age gaps and
402 health status. This suggests that biological age gaps may not have a consistent interpretation across
403 age groups, health-status groups, populations, or data-acquisition settings. Future studies should
404 therefore move beyond improving prediction accuracy within a single development cohort and evaluate
405 whether biological age gaps retain robust and clinically meaningful interpretations across diverse
406 settings. Downstream association studies and clinical models should explicitly account for age-
407 dependent associations rather than assuming a constant relationship between biological age gap and
408 health status.

409
410 This study also has several limitations. First, the “healthy” participants included here cannot be
411 considered completely healthy; rather, they were defined as relatively healthy based on the absence of
412 recorded disease categories (details introduced in the Method section, where we excluded a broad
413 range of diseases). Definitions of health vary substantially across studies and often depend on available
414 disease records and diagnostic criteria. Second, we used logistic regression to analyse the association
415 between RAG and disease as a representative example of clinical analysis. Future work could
416 incorporate additional modelling frameworks, such as Cox regression for longitudinal risk prediction
417 [15,22], to provide a broader evaluation of clinical utility. Third, external validation was limited to three
418 cohorts from the UK and Brazil, with a restricted range of imaging devices and population
419 characteristics. To improve broader applicability, future studies would include more geographically
420 diverse cohorts, wider demographic representation and additional imaging devices, enabling a more
421 comprehensive evaluation of retinal age model generalisability.

422
423 In conclusion, this study investigated the fundamental characteristics of retinal age models and provided
424 insights for their more reliable interpretation and applications. We identified inconsistent RAG and health
425 associations across age subgroups, driven by differential RTM effects between healthy and unhealthy
426 cohorts, and observed varied model generalisability across external cohorts and intended applications.
427 These findings highlight the need for age-dependent RAG interpretation and application-specific
428 external evaluation for reliable clinical applications. More broadly, this work highlights the importance of
429 systematically investigating RTM and differential RTM effects across subgroups, in order to establish
430 more reliable interpretations and appropriate clinical applications for biological ageing markers.

431 **Methods**

432 **Model description**

433 The architecture of the retinal age model is illustrated in [Supplementary Fig. 12](#). The age estimation was
434 formulated as a regression problem, similar to previous studies [15,17,20]. Specifically, we employed
435 the DINOv3-Large (vit_large_patch16_dinov3.lvd1689m) [59] as the primary backbone for feature
436 extraction, utilising the standard Vision Transformer (ViT) [60] architecture initialised with weights
437 pretrained on the LVD142M dataset. The DINOv3 model processed input 224×224 images with a patch
438 size of 16. Patch tokens (excluding the prefix and class tokens) were averaged along the sequence
439 dimension and layer-normalised, yielding a 1,024-dimensional feature vector. A custom regression head
440 consists of a linear layer reducing the dimension to 32, a ReLU activation, a dropout layer with dropout
441 probability 0.5, and a final linear transformation to produce a single scalar output (e.g., the estimated
442 retinal age).

443 **Model training**

444 All backbone and head parameters were jointly fine-tuned. The chronological-age target was normalised
445 by subtracting 50 and dividing by 50; all reported metrics are computed after rescaling predictions back
446 to the original years. Data augmentation followed RETFound [41]: random resized crop to 224×224,
447 random horizontal flipping, AutoAugment (rand-m9-mstd0.5-inc1), random erasing with probability 0.25,
448 and ImageNet-mean / ImageNet-std normalisation. Optimisation used AdamW with weight decay 0.05,
449 layer-wise learning-rate decay of 0.65, gradient clipping at 1.0 and mixed-precision training. Training
450 was performed with a batch size of 32 for 50 epochs. The learning rate was warmed up linearly from 0
451 to 6×10^{-4} over the first 10 epochs, followed by a cosine annealing schedule that decayed the rate from
452 6×10^{-4} to 1×10^{-6} over the remaining 40 epochs. After each epoch, model performance was evaluated
453 on the retinal-age validation set; the checkpoint with the lowest mean absolute error (MAE) was retained
454 for downstream analysis.

455 **Datasets**

456 The retinal age model was trained on the AlzEye [48] cohort. Internal evaluation used a held-out AlzEye
457 internal test set, and external generalisation was assessed on three independent cohorts: UK Biobank
458 [49], BRSET [50], and mBRSET [51]. The AlzEye dataset is a comprehensive retrospective cohort
459 linking retinal imaging to systemic clinical records for over 200,000 patients. It was acquired at
460 Moorfields Eye Hospital (London, UK) predominantly using Topcon systems ($\approx 99\%$, with the remainder
461 using Carl Zeiss Meditec). The UK Biobank is a large prospective epidemiological study encompassing
462 approximately half a million participants, which includes a subset of over 80,000 individuals with retinal
463 imaging. This cohort predominantly consists of White British participants (approximately 90%), and the
464 retinal images were acquired using the Topcon 3D OCT-1000. BRSET offers substantial geographic and
465 ethnic diversity from a South American Brazilian cohort of over 16,000 patients, featuring a demographic
466 mix including Pardo (mixed race), White, and Black individuals, with images collected via desktop
467 Canon and Nikon devices. mBRSET expands on this diversity by utilising the handheld Phelcom Eyer
468 portable camera, introducing unique mobile imaging characteristics on a similarly diverse population. A

summary of detailed device information and demographic statistics for all datasets are provided in [Supplementary Tab. 1](#).

Quality control

For all four datasets, we utilised AutoMorph [61] to classify images as Good, Usable, or Reject. Only images in the Good and Usable images were selected for further processing, as shown in [Supplementary Fig. 1,4,5](#).

Disease labelling

Each subject was labelled as "unhealthy" if at least one positive disease record was present, otherwise "healthy."

For the AlzEye dataset, disease labels were obtained through linked hospital records using International Classification of Diseases, 10th Revision (ICD-10) codes [62]. These labels were organised into four categories: cardiovascular, chronic systemic, ocular, and neurodegenerative diseases. Cardiovascular diseases included atrial fibrillation, heart failure, myocardial infarction, intracerebral haemorrhage, ischaemic stroke, aortic stenosis, aortic regurgitation, mitral stenosis, and mitral regurgitation. Chronic systemic diseases included diabetes mellitus and hypertension. Ocular diseases included glaucoma, diabetic retinopathy, and age-related macular degeneration. Neurodegenerative diseases included dementia and Parkinson's disease. The full ICD-10 code definitions are provided in [Supplementary Tab. 13](#). Neurodegenerative diseases were included to define the healthy and overall unhealthy groups, but were not included in disease subgroup analyses because of limited sample sizes. Throughout this work, we refer collectively to the overall unhealthy group together with the three analysed disease subcategories, cardiovascular, chronic systemic, and ocular diseases, as the four "disease groups".

For the UK Biobank [40] dataset, disease information was similarly obtained using ICD-10 [62] and self-report codes (UKB Data-Field 20002). The disease labels included atrial fibrillation, heart failure, myocardial infarction, intracerebral haemorrhage, ischaemic stroke, aortic stenosis, aortic regurgitation, mitral stenosis, mitral regurgitation, diabetes mellitus, hypertension, dementia, Parkinson's disease, glaucoma, cataract, and diabetic retinopathy. The full ICD-10 and self-report code definitions are provided in [Supplementary Tab. 13](#). For BRSET [50] and mBRSET [51], disease labels were provided in the metadata of the original datasets. For BRSET, the disease labels are diabetic retinopathy, macular oedema, scar, nevus, age-related macular degeneration, vascular occlusion, hypertensive retinopathy, drusen, haemorrhage, retinal detachment, myopic fundus, increased cup-disc ratio, and other abnormalities. For mBRSET, the disease labels are systemic hypertension, obesity, vascular disease, acute myocardial infarction, nephropathy, neuropathy, and diabetic foot.

Dataset splits

AlzEye patients were partitioned at the patient level in two stages. As illustrated in [Supplementary Fig. 1](#), healthy patients were first randomly split 50%/50%; one half (12,306 patients / 67,718 images) was further split at the patient level into 80% retinal-age training and 20% retinal-age validation subsets. The remaining healthy patients together with all unhealthy patients formed a combined set (48,559 patients / 260,046 images), which was split 75%/25% at the patient level into a downstream training subset

(36,418 patients / 194,686 images) and an internal test subset (12,141 patients / 65,360 images). Each external cohort was randomly split 50%/50% at the patient level into training and test subsets. Per-set patient and image counts together with chronological-age distributions are shown in [Supplementary Figs. 1-8](#).

Quantitative evaluation metrics for retinal age models

To quantitatively evaluate the retinal age estimation model, we utilised three primary metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the squared Pearson correlation coefficient (Pearson's r^2) between estimated retinal age and chronological age. MAE and RMSE served as standard measures of regression accuracy between the estimated retinal age and chronological age, where lower metric values indicate better estimation performance. Specifically, MAE reflects the average magnitude of prediction errors, while RMSE provides sensitivity to outliers by penalising larger deviations more heavily. Pearson's r^2 quantified the strength of the linear correlation between estimated retinal age and chronological age, with higher values indicating stronger linear agreement.

Association study

In internal evaluation, we performed a logistic regression with the disease label (healthy vs. disease) as the dependent variable and RAG as the primary independent variable of interest, adjusted for chronological age as a covariate, separately for each of the four disease groups versus healthy. The strength of the association was quantified using odds ratios (OR) per +1 year increase in RAG with corresponding 95% CIs, and statistical significance was determined by the p-value calculated using the Wald test.

To account for the differential RTM and examine how this association varies with chronological age, we extended the logistic regression by additionally including an Age $\times$ RAG interaction term, with chronological age, RAG, and their interaction as predictors. The implied OR of disease per +1 year increase in RAG was then computed as a function of chronological age, with 95% CIs from a patient-level cluster bootstrap. The same analysis was applied on the AlzEye internal test set and on each external cohort's test set after refitting the logistic regression on the cohort's training set.

Linear regression of RAG on chronological age

For each group (healthy or one of the four disease groups), we fit a single ordinary least squares regression on the combined data of healthy and disease groups:

$$RAG = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Group + \beta_3 \cdot Age \times Group + \varepsilon,$$

with "Group" coded 0 for healthy and 1 for the disease group, and Age mean-centred (original age minus mean age) so that β_0 and β_2 correspond to the mean age. The healthy slope (β_1) and the disease-group slope ($\beta_1 + \beta_3$) were reported with 95% confidence intervals. The Wald p-value on the interaction coefficient β_3 was reported as the test for slope difference between healthy and the disease group.

542 **Age calibration of RAG**

543 Two age-calibrators were fit on the healthy subset of the downstream training set by regressing RAG on
544 chronological age: a linear calibrator

545
$$\text{Calibrator}_{\text{Linear}}(\text{Age}) = \alpha_l + \beta_l \cdot \text{Age},$$

546 and a quadratic calibrator

547
$$\text{Calibrator}_{\text{quadratic}}(\text{Age}) = \alpha_q + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Age}^2.$$

548 The calibrated RAG was then defined as

549
$$\text{RAG}_{\text{cal}} = \text{RAG} - \text{Calibrator}(\text{Age}).$$

550 Within-group slopes and the interaction p-value were re-computed on the internal test set with the
551 calibrated RAG in place of the raw RAG.

552 **Balanced training**

553 To assess whether the differential RTM was driven by the non-uniform age distribution of the retinal-age
554 training set, we trained the retinal age model from scratch on a uniform-age training distribution. Each
555 patient's sampling weight was set proportional to $1 / \max(N_{\text{age-bin}}, N_{\text{min}})$ with 5-year age bins and
556 $N_{\text{min}} = 50$ to avoid extreme upweighting of sparse bins. All other training hyperparameters were
557 unchanged. The balanced model was evaluated on the same internal test set and within-group RAG-
558 age slopes were re-computed.

559 **Disease classification**

560 For disease classification, we considered three logistic regression models:

- 561 - $\text{logit}(\text{disease}) \sim \text{Age}$
562 - $\text{logit}(\text{disease}) \sim \text{Age} + \text{RAG}$
563 - $\text{logit}(\text{disease}) \sim \text{Age} + \text{RAG} + \text{Age} \times \text{RAG}$

564 All three models were fitted on the AlzEye downstream training set, separately for overall unhealthy and
565 for each of the three disease categories versus healthy. Discrimination on the AlzEye internal test set
566 was reported as the area under the receiver operating characteristic curve (AUC). For threshold-based
567 metrics (sensitivity, specificity, and F1 score), the operating threshold was selected by the Youden
568 criterion ($J = \text{sensitivity} + \text{specificity} - 1$) on the AlzEye downstream training set.

569 For external generalisation, each external cohort was evaluated on its test set in two ways: off-the-shelf
570 and externally calibrated. Under the off-the-shelf evaluation, the AlzEye-fit logistic regression
571 coefficients and the AlzEye Youden threshold were applied directly to the external test predictions,
572 without any cohort-specific adjustment. Under the externally-calibrated evaluation, the logistic
573 regression was refitted on the external cohort's training set, and the operating threshold was re-derived
574 as the cohort-specific Youden point on that training set. The retinal-age model was not re-trained in
575 either evaluation.

576 **Ethic approval**

577 Ethical approval for this study was obtained from the London-Central Research Ethics Committee
578 (18/LO/1163, approved 1 August 2018), Confidential Advisory Group for Section 251 support
579 (18/CAG/0111, approved 13 September 2018). The National Health Service Health Research Authority
580 gave final approval on 13 September 2018. Moorfields Eye Hospital (MEH) NHS Foundation Trust
581 validated the de-identifications for MEH data. Only de-identified retrospective data were used for
582 research. The requirement for informed consent using AlzEye data was waived following section 251
583 supported by the UK CAG.

584 **Data availability**

585 The MEH data consists of routinely collected healthcare data. Owing to their sensitive nature, the
586 dataset is subject to controlled access by means of a structured application process. The AlzEye
587 dataset is subject to the contractual restrictions of the data sharing agreements between National Health
588 Service England, Moorfields Eye Hospital and University College London due to data privacy and ethical
589 requirements. Interested collaborators to MEH data should contact the chief investigator P.A.K. at
590 p.keane@ucl.ac.uk and submit a formal submission at <https://www.insight.hdrhub.org/insight-data>. The
591 data Trust Advisory Board will then review all the requests and grants. A formal data transfer agreement
592 will be required upon approval, and the cost will be generated. Generally, all these requests for access
593 to the data will be responded to within 3 months.

594 This research has been conducted using the UK Biobank Resource under application number 36741.
595 The UK Biobank data are available for approved researchers through the UK Biobank Access
596 Management System (<https://www.ukbiobank.ac.uk>). Publicly available datasets used in this study
597 include mBRSET, accessible at <https://physionet.org/content/mbrset/1.0/>, and BRSET, accessible at
598 <https://physionet.org/content/brazilian-ophthalmological/1.0.1/>.

599 **Code availability**

600 The code for the development of this study is available at our GitHub repository at
601 <https://github.com/HORIZONHealthcare/RAG-RTM>. We implemented the network design, model
602 training, and model evaluation using PyTorch (<https://pytorch.org>) v2.2.0. The DINOv3 model weights
603 are available at <https://github.com/facebookresearch/dinov3>. Images were processed with the
604 automated retinal image analysis tool AutoMorph v.1.0 (<https://github.com/rmaphoh/AutoMorph>).
605 Primary Python (v3.10.18) packages used in this study include Seaborn v0.13.2, NumPy v1.26.4,
606 Matplotlib v3.10.7, Pandas v2.3.2, SciPy v1.15.3, and Scikit-Image v0.25.2.

607 **Acknowledgement**

608 Y.Z. is supported by Wellcome Trust Early-career Award (318987/Z/24/Z) and Moorfields Eye Charity
609 Equipment Grant (EQR-25B-102). Z.Z. is supported by the National Health and Medical Research
610 Council (NHMRC) Investigator Grant (Emerging Leadership 1, APP2010072), NHMRC Investigator
611 Grant (Emerging Leadership 2, APP2041559) and Medical Research Future Fund (MRFF) Early to Mid-
612 Career Researchers Grant (APP2027411). Y.C.T. is supported by the National Medical Research
613 Council of Singapore (NMRC/MOH/ HCSAINV21nov-0001). C.Y.C. is supported by the Research

614 Grants Council Hong Kong (General Research Fund, ref.14101324). D.C.A. is supported by the
615 Engineering and Physical Sciences Research Council (EP/M020533/1, EP/R014019/1 and
616 EP/V034537/1). P.A.K. is supported by a UK Research & Innovation Future Leaders Fellowship
617 (MR/T019050/1), the Moorfields Eye Charity with The Rubin Foundation Charitable Trust (GR001753),
618 and an Alcon Research Institute Senior Investigator Award. We acknowledge the computational
619 resources supported by the UCL Computer Science Cluster and UCL Advanced Research Computing
620 UAI platform.

621

622 **Reference**

623 1. L'opez-Ot'in C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell*.
624 2013;153: 1194–1217.

625 2. Alber J, Bernoth M, Bennett S, Hill S. Aging with artificial intelligence. *GeroScience*. 2019;41: 733–
626 741.

627 3. Belsky DW, Caspi A, Houts R, Cohen HJ, Corcoran DL, Danese A, et al. Quantification of biological
628 aging in young adults. *Proceedings of the National Academy of Sciences*. 2015;112: E4104–E4110.

629 4. Beltrán-Sánchez H, Palloni A, Huangfu Y, McEniry MC. Modeling biological age and its link with the
630 aging process. *PNAS Nexus*. 2022;1: gac135.

631 5. MacDonald ME, Pike GB. MRI of healthy brain aging: A review. *NMR Biomed*. 2021;34: e4564.

632 6. de Lange A-MG, Anatürk M, Suri S, Kaufmann T, Cole JH, Griffanti L, et al. Multimodal brain-age
633 prediction and cardiovascular risk: The Whitehall II MRI sub-study. *Neuroimage*. 2020;222: 117292.

634 7. Cole JH, Poudel GR, Tsagkrasoulis D, Caan MWA, Steves C, Spector TD, et al. Predicting brain
635 age with deep learning from raw imaging data results in a reliable and heritable biomarker.
636 *Neuroimage*. 2017;163: 115–124.

637 8. Cole JH, Franke K. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers.
638 *Trends Neurosci*. 2017;40: 681–690.

639 9. Lee J, Burkett BJ, Min H-K, Senjem ML, Lundt ES, Botha H, et al. Deep learning-based brain age
640 prediction in normal aging and dementia. *Nat Aging*. 2022;2: 412–424.

641 10. Bontempi D, Zalay O, Bitterman DS, Birkbak N, Shyr D, Haugg F, et al. FaceAge, a deep learning
642 system to estimate biological age from face photographs to improve prognostication: a model
643 development and validation study. *Lancet Digit Health*. 2025;7: 100870.

644 11. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14: 3156.

645 12. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sada S, et al. Genome-wide methylation
646 profiles reveal quantitative views of human aging rates. *Mol Cell*. 2013;49: 359–367.

647 13. Mavrommatis C, Belsky DW, Ying K, Moqri M, Campbell A, Richmond A, et al. An unbiased
648 comparison of 14 epigenetic clocks in relation to 174 incident disease outcomes. *Nat Commun*.
649 2025;16: 11164.

650 14. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing.
651 *Nat Rev Genet*. 2018;19: 371–384.

652 15. Zhu Z, Shi D, Peng G, Tan Z, Shang X, Hu W, et al. Retinal age gap as a predictive biomarker for
653 mortality risk. *Br J Ophthalmol*. 2023;107: 547–554.

654 16. Wang Y, Xiao S, Liu B, Jiang R, Liu Y, Hang Y, et al. Organ-specific proteomic aging clocks predict
655 disease and longevity across diverse populations. *Nat Aging*. 2026;6: 162–180.

656 17. Yu Z, Chen R, Gui P, Wang W, Razzak I, Alinejad-Rokny H, et al. A cross population study of
657 retinal aging biomarkers with longitudinal pre-training and label distribution learning. *NPJ Digit Med*.
658 2025;8: 344.

659 18. Ferrucci L, Barzilai N, Belsky DW, Gladyshev VN. How to measure biological aging in humans. *Nat*
660 *Med*. 2025;31: 1057.

661 19. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of
662 cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*.

2018;2: 158–164.

20. Nusinovici S, Rim TH, Yu M, Ong GSW, Tan GSW, Lee CL, et al. Retinal age gap as a predictive biomarker of future cardiovascular disease events. *Age Ageing*. 2022;51: afac115.
21. Zhu Z, Hu W, Chen R, Xiong R, Wang W, Shang X, et al. Retinal age gap as a predictive biomarker of stroke risk. *BMC Med*. 2022;20: 466.
22. Hu W, Wang W, Wang Y, Chen Y, Shang X, Liao H, et al. Retinal age gap as a predictive biomarker of future risk of Parkinson's disease. *Age Ageing*. 2022;51: afac062.
23. Zhu Z, Chen Y, Wang W, Wang Y, Hu W, Shang X, et al. Association of retinal age gap with arterial stiffness and incident cardiovascular disease. *Stroke*. 2022;53: 3320–3328.
24. Nusinovici S, Rim TH, Li H, Yu M, Deshmukh M, Quek TC, et al. Application of a deep-learning marker for morbidity and mortality prediction derived from retinal photographs: a cohort development and validation study. *Lancet Healthy Longev*. 2024;5: 100593.
25. Zhu Z, Liu D, Chen R, Hu W, Liao H, Kiburg K, et al. The association of retinal age gap with metabolic syndrome and inflammation. *J Diabetes*. 2023;15: 237–245.
26. Flammer J, Konieczka K, Bruno RM, Virdis A, Flammer AJ, Taddei S. The eye and the heart. *Eur Heart J*. 2013;34: 1270–1278.
27. London A, Benhar I, Schwartz M. The retina as a window to the brain—from eye research to CNS disorders. *Nat Rev Neurol*. 2013;9: 44–53.
28. Wagner SK, Fu DJ, Faes L, Liu X, Huemer J, Khalid H, et al. Insights into systemic disease through retinal imaging-based oculomics. *Transl Vis Sci Technol*. 2020;9: 6–6.
29. Zhu Z, Wang Y, Qi Z, Hu W, Zhang X, Wagner SK, et al. Oculomics: Current concepts and evidence. *Prog Retin Eye Res*. 2025; 101350.
30. Lee J, Han M, Wang K, Butler LR, Sinclair DA. Epigenetic Reprogramming for Ocular Aging and Disease: Mechanisms, Biomarkers, and the Road to the Clinic. *Prog Retin Eye Res*. 2026; 101442.
31. Hu W, Wang W, Wang Y, Chen Y, Shang X, Liao H, et al. Retinal age gap as a predictive biomarker of future risk of Parkinson's disease. *Age Ageing*. 2022;51: afac062.
32. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, et al. The transcriptional landscape of age in human peripheral blood. *Nat Commun*. 2015;6: 8570.
33. Lehallier B, Gate D, Schaum N, Nanasi T, Lee SE, Yousef H, et al. Undulating changes in human plasma proteome profiles across the lifespan. *Nat Med*. 2019;25: 1843–1850.
34. Ahadi S, Zhou W, Schüssler-Fiorenza Rose SM, Sailani MR, Contrepois K, Avina M, et al. Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nat Med*. 2020;26: 83–90.
35. Rim TH, Lee CL, Tham Y-C, Cheung CY, Yu M, Ang M, et al. Prediction of biological age and risk of cardiovascular disease from retinal fundus images using deep learning. *Lancet Digit Health*. 2020;2: e513–e524.
36. Nielsen C, Stanley EAM, Wilms M, Forkert ND. Assessment of demographic bias in retinal age prediction machine learning models. *Front Artif Intell*. 2025;8: 1653153.
37. Zhang Y, Shang X, Zhu Z, Zhang X, Huang Y, Ge Z, et al. Retinal age gap as a predictive biomarker for risk of kidney failure: a population-based cohort study. *Am J Kidney Dis*. 2022;80: 718–726.
38. Antaki F, Kerexeta-Sarriegi J, Reis APR, Zhu Z, Chen R, Hu W, et al. The association of retinal age

gap with schizophrenia: a cross-sectional analysis. *Schizophr Res.* 2025;283: 180–187.

39. Chen R, Zeng X, Hu W, Jeyarajan D, Yu Z, Wang W, et al. Accelerated retinal ageing and multimorbidity in middle-aged and older adults. *GeroScience.* 2025;47: 4291–4300.

40. Wang J, Gao Y, Wang F, Zeng S, Li J, Miao H, et al. Accurate estimation of biological age and its application in disease prediction using a multimodal image Transformer system. *Proc Natl Acad Sci U S A.* 2024;121: e2308812120.

41. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, et al. A foundation model for generalizable disease detection from retinal images. *Nature.* 2023;622: 156–163.

42. Nielsen C, Wilms M, Forkert ND. A novel foundation model-based framework for multimodal retinal age prediction. *IEEE J Transl Eng Health Med.* 2025;13: 299–309.

43. Nielsen C, Souza R, Wilms M, Forkert ND. Foundation model-driven distributed learning for enhanced retinal age prediction. *J Am Med Inform Assoc.* 2024;31: 2550–2559.

44. Barnett AG, Van Der Pols JC, Dobson AJ. Regression to the mean: what it is and how to deal with it. *Int J Epidemiol.* 2005;34: 215–220.

45. Liang H, Zhang F, Niu X. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. *Hum Brain Mapp.* 2019;40: 3143–3152.

46. Smith SM, Vidaurre D, Alfaro-Almagro F, Nichols TE, Miller KL. Estimation of brain age delta from brain imaging. *bioRxiv.* bioRxiv; 2019. doi:[10.1101/560151](https://doi.org/10.1101/560151)

47. Wen J. Refining the generation, interpretation and application of multi-organ, multi-omics biological aging clocks. *Nat Aging.* 2025;5: 1897–1913.

48. Wagner SK, Hughes F, Cortina-Borja M, Pontikos N, Struyven R, Liu X, et al. AlzEye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353 157 patients in London, UK. *BMJ Open.* 2022;12: e058552.

49. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015;12: e1001779.

50. Nakayama LF, Restrepo D, Matos J, Ribeiro LZ, Malerbi FK, Celi LA, et al. BRSET: A Brazilian Multilabel Ophthalmological Dataset of retina fundus photos. *PLOS Digit Health.* 2024;3: e0000454.

51. Wu C, Restrepo D, Nakayama LF, Ribeiro LZ, Shuai Z, Barboza NS, et al. MBRSET: A portable retina fundus photos benchmark dataset for clinical and demographic prediction. *medRxiv.* 2024. doi:[10.1101/2024.07.11.24310293](https://doi.org/10.1101/2024.07.11.24310293)

52. Oh HS-H, Rutledge J, Nachun D, Pálovics R, Abiose O, Moran-Losada P, et al. Organ aging signatures in the plasma proteome track health and disease. *Nature.* 2023;624: 164–172.

53. Zalesky A, Wen J, Tian YE. From whole-body to organ-specific biological age clocks. *Nat Aging.* 2026;6: 961–969.

54. Ying K, Paulson S, Eames A, Tyshkovskiy A, Li S, Eynon N, et al. A unified framework for systematic curation and evaluation of aging biomarkers. *Nat Aging.* 2025;5: 2323–2339.

55. Shah J, Siddiquee MMR, Su Y, Wu T, Li B. Ordinal classification with distance regularization for Robust brain age prediction. *IEEE Winter Conf Appl Comput Vis.* 2024;2024: 7867–7876.

56. El Khoury LY, Gorrie-Stone T, Smart M, Hughes A, Bao Y, Andrayas A, et al. Systematic underestimation of the epigenetic clock and age acceleration in older subjects. *Genome Biol.* 2019;20: 283.

746 57. Confounds in multivariate differential expression analyses of newly acquired datasets: A Case
747 Study on Brain Age.

748 58. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, et al. DNA methylation aging clocks:
749 challenges and recommendations. *Genome Biol.* 2019;20: 249.

750 59. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: Learning robust
751 visual features without supervision. *arXiv [cs.CV]*. 2023. Available: <http://arxiv.org/abs/2304.07193>

752 60. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is
753 worth 16x16 words: Transformers for image recognition at scale. *arXiv [cs.CV]*. 2020. Available:
754 <http://arxiv.org/abs/2010.11929>

755 61. Zhou Y, Wagner SK, Chia MA, Zhao A, Woodward-Court P, Xu M, et al. AutoMorph: Automated
756 retinal vascular morphology quantification via a deep learning pipeline. *Transl Vis Sci Technol.*
757 2022;11: 12.

758 62. International Statistical Classification of Diseases and related health problems: Alphabetical index.
759 2004;3. Available:
760 https://books.google.com/books?hl=en&lr=&id=Tw5eAtsatiUC&oi=fnd&pg=PA1&dq=International+statistical+classification+of+diseases+and+related+health+problems&ots=o5eXm_qKhN&sig=qGWV8oI2w5iPSJ8s6BejFsyzzYE
761
762

763

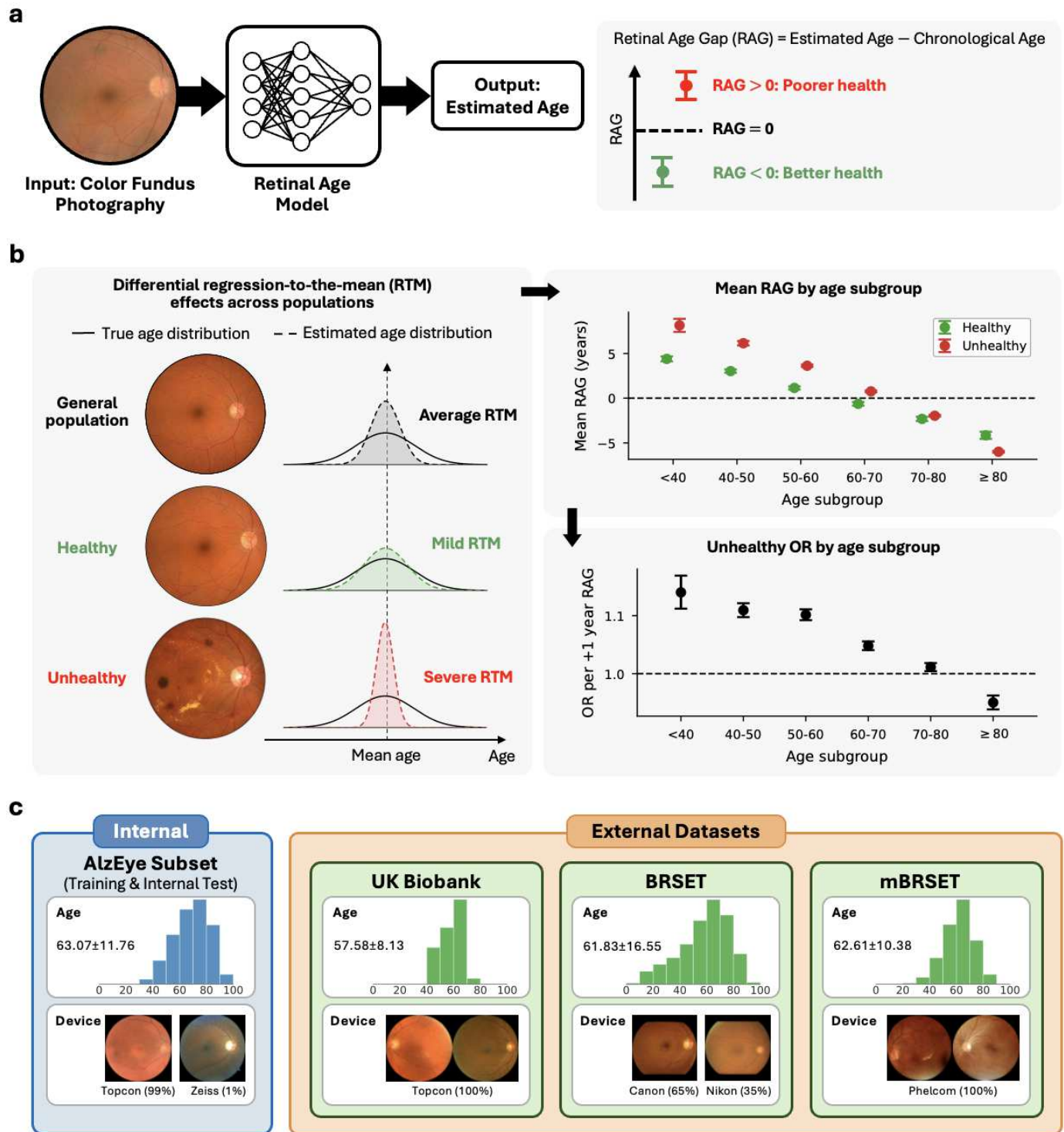


Fig. 1. Study overview and differential regression-to-the-mean (RTM) effects in retinal age models. **a.** Retinal age estimation and retinal age gap (RAG) calculation. A retinal age model estimates retinal age from colour fundus photography. Retinal age estimation and retinal age gap (RAG) calculation. A retinal age model estimates retinal age from colour fundus photography. RAG is defined as estimated age minus chronological age. It is commonly interpreted as an ageing-related biomarker, where an older estimated retinal age indicates poorer health, whereas a younger estimated retinal age indicates slower ageing and healthier status. **b.** Differential RTM effects and their consequences for RAG interpretation. Left: schematic illustration of RTM across populations. The general population exhibits average RTM, the healthy population shows mild RTM, and the unhealthy population shows severe RTM, with estimated ages

more strongly concentrated around the mean age. Top right: differential RTM leads to distinct age-varying mean RAG patterns in healthy and unhealthy groups. RAG decreases with age in both groups, but declines more steeply in the unhealthy group, reducing or reversing the RAG difference at older ages. Bottom right: these age-varying RAG patterns induce an age-dependent association between RAG and unhealthy conditions. The odds ratio (OR) for unhealthy conditions per 1-year increase in RAG attenuates with age and may reverse in the oldest age subgroup. Full results for RAG and OR are reported in [Supplementary Tab. 2](#). **c.** Datasets used in this study: The retinal age model was developed and internally tested on the AlzEye dataset. External validation was performed on three independent datasets: UK Biobank, BRSET, and mBRSET.

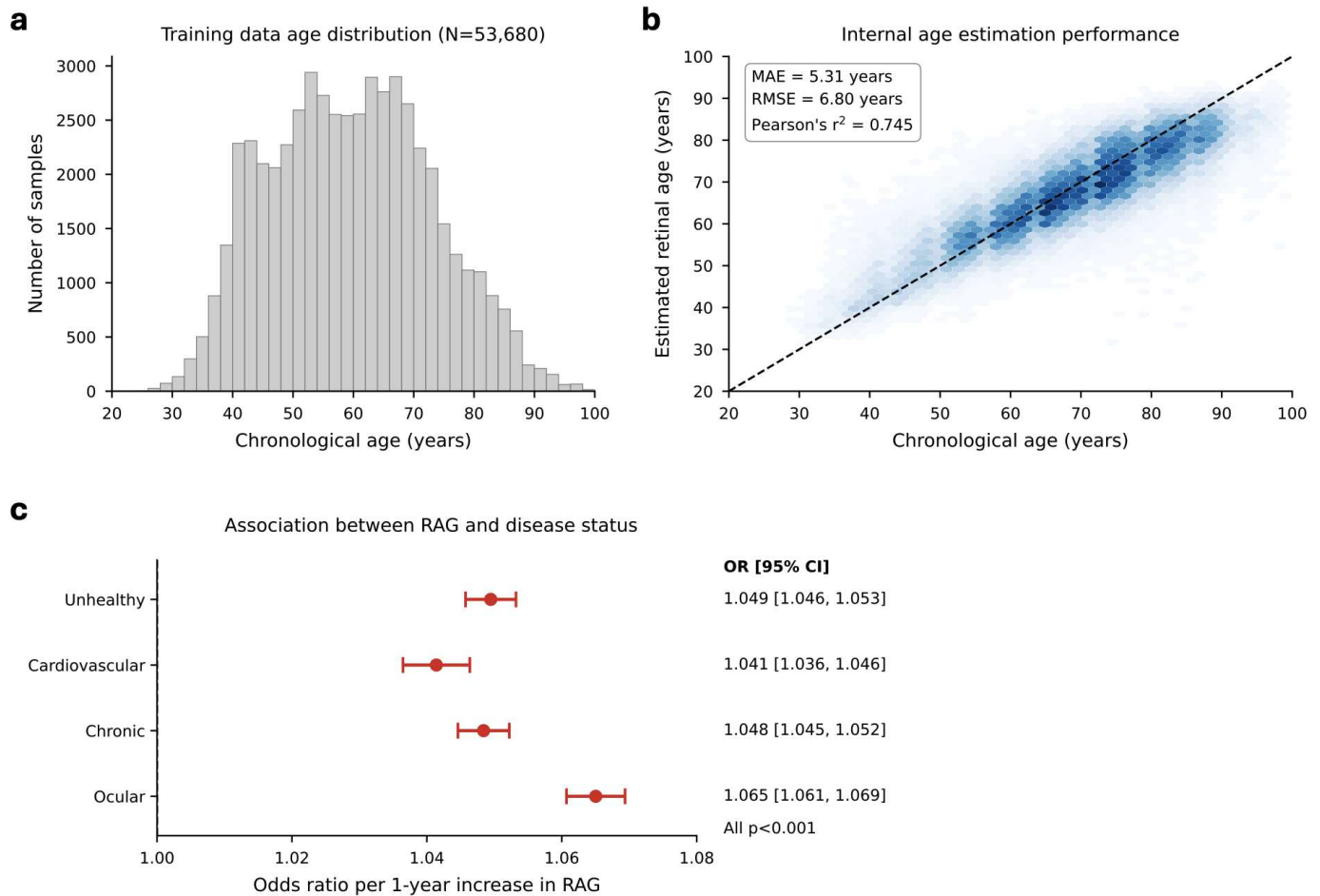


Fig. 2. Internal retinal age estimation and association with disease on the AlzEye internal test set. a.

Chronological age distribution of the retinal-age training set used to train the retinal age model. **b.** Internal retinal age model performance. Hexbin density of estimated retinal age vs chronological age on the internal test set. The dashed line denotes the identity line (estimated = chronological). Annotated metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the squared Pearson correlation coefficient (Pearson's r^2) between estimated and chronological age. Lower MAE and RMSE indicate smaller estimation error. Higher Pearson's r^2 indicates a stronger linear correlation between estimated and chronological age. **c.** Association between RAG and disease conditions. Logistic regression was performed with RAG as the independent variable and each disease condition as the outcome, adjusted for chronological age. Reported are odds ratios (OR) per 1-year increase in RAG with 95% confidence intervals; p-values are from the Wald test. All four disease conditions show significant positive associations with RAG (all $p < 0.001$). Full results are reported in [Supplementary Tab. 3-4](#).

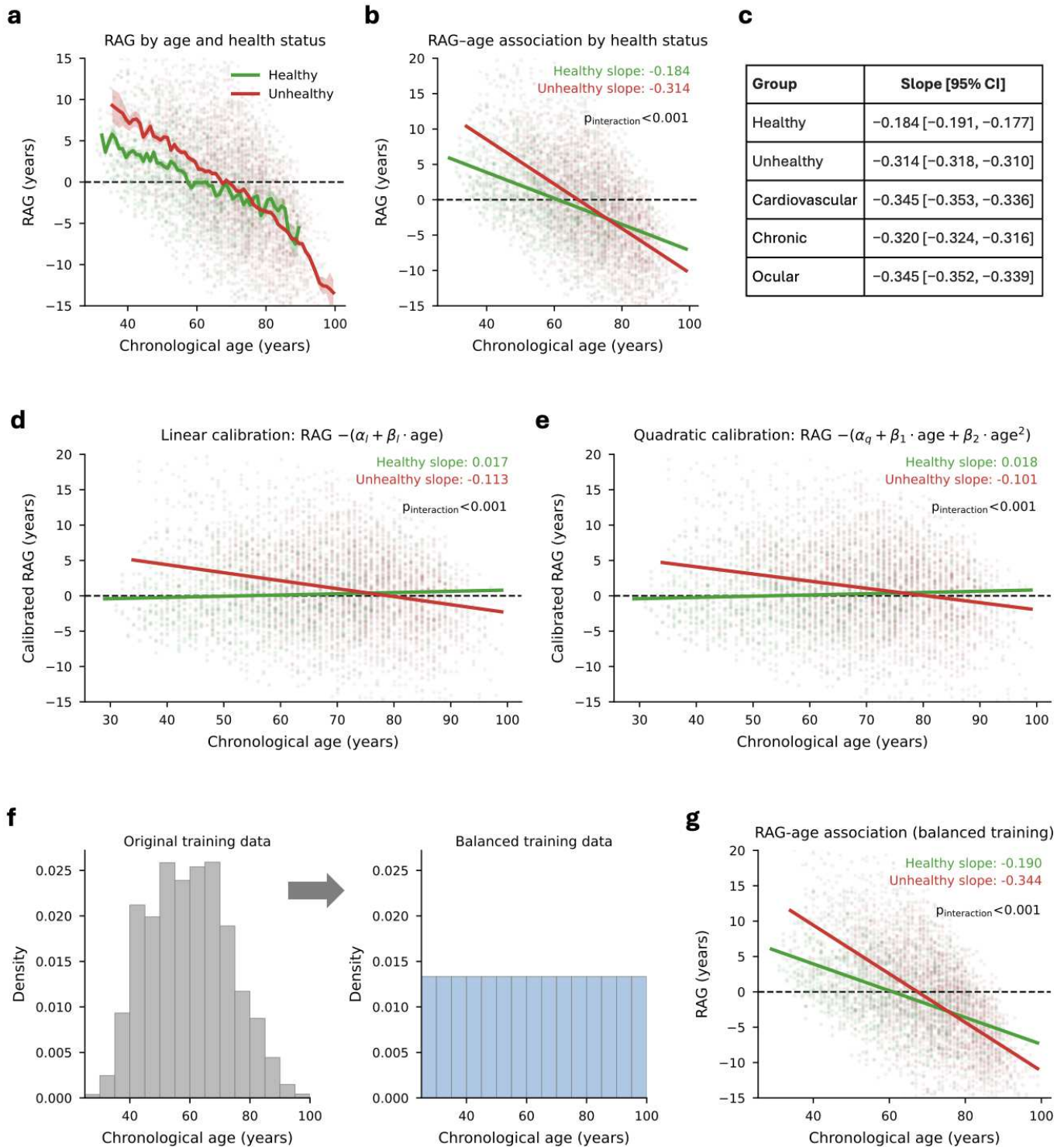


Fig. 3. Differential RTM persists after RAG calibration and training-distribution balancing. **a-c.** Differential RTM effects between healthy and unhealthy groups on the AlzEye internal test set. **a.** Mean RAG vs. chronological age stratified by health status. Shaded bands are 95% CIs from a patient-level cluster bootstrap. **b.** Linear regression of RAG on chronological age, group (0: healthy; 1: unhealthy), and their interaction. The unhealthy group shows a steeper negative slope than the healthy group (interaction $p < 0.001$ from Wald test on Age $\times$ Group term), indicating stronger RTM. **c.** RAG-age Slope estimates with 95% CIs for healthy and four disease groups. All disease groups show steeper negative slopes than the healthy group, indicating that differential RTM is a general phenomenon across disease groups. **d-e.** Linear and quadratic RAG calibration based on the healthy subset of the downstream training

set. Calibrated RAG was obtained by removing the expected age-related RAG trend estimated from a linear RAG-age calibration model (**d**) or a quadratic RAG-age calibration model (**e**). Calibration largely removes the RAG-age trend in the healthy group. However, a significant negative slope persists in the unhealthy group under both calibrations, with significant Age $\times$ Group interactions (interaction $p < 0.001$, Wald test). **f-g**. Training-distribution rebalancing. **f**. Original training-age distribution (left) and the uniform-age distribution obtained by inverse-frequency resampling (right). **g**. RAG-age association after retraining the retinal age model on the balanced training distribution. The unhealthy group still shows a steeper negative RAG-age slope than the healthy group, with a significant Age $\times$ Group interaction (interaction $p < 0.001$, Wald test), indicating that training-age balancing does not resolve differential RTM. Full results are reported in [Supplementary Tab. 5-7](#).

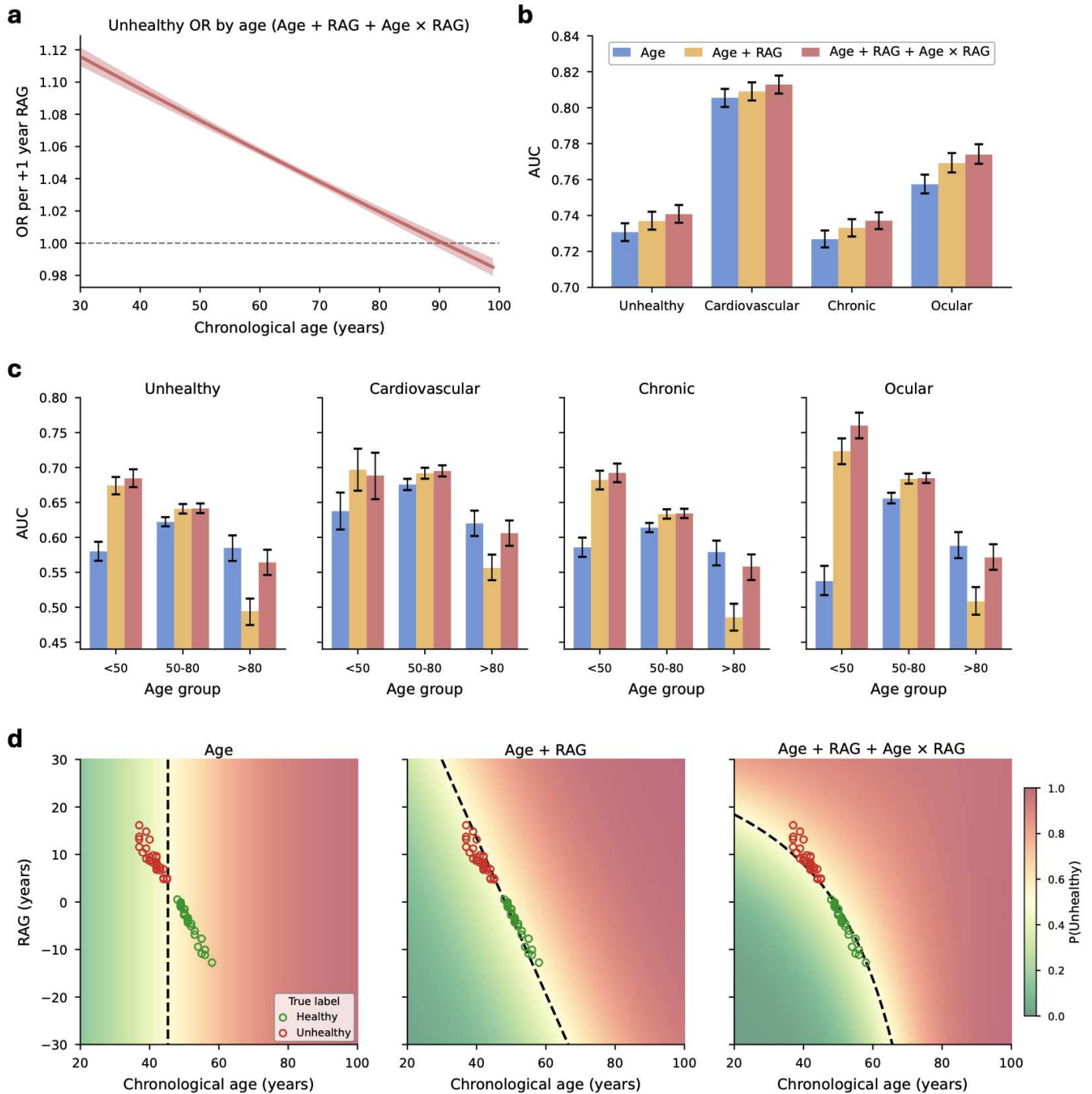


Fig. 4. Incorporating RAG into downstream disease classification under RTM. **a.** Age-specific odds ratio for unhealthy status per 1-year increase in RAG, estimated from the Age + RAG + Age $\times$ RAG logistic model. The declining odds ratio shows that the association between RAG and unhealthy status is age-dependent under RTM. Shaded band denotes the 95% CI from patient-level cluster bootstrap. **b.** Overall AUC for unhealthy, cardiovascular, chronic and ocular disease prediction on the AlzEye internal test set using three logistic regression models trained on the AlzEye downstream training set. Error bars denote 95% CI. **c.** AUC within age subgroups for the same outcomes and models as in **b**, showing that the added value of RAG differs across chronological age. **d.** Probability plots in the Age-RAG plane for the unhealthy outcome under the three logistic regression models, showing predicted $P(\text{unhealthy})$ and the corresponding decision boundaries. Highlighted points denote internal test cases that are misclassified by the Age-only and Age + RAG models but correctly classified by the Age + RAG + Age $\times$ RAG model. Full results are reported in Tab. Full results are reported in [Supplementary Tab. 8-9](#).

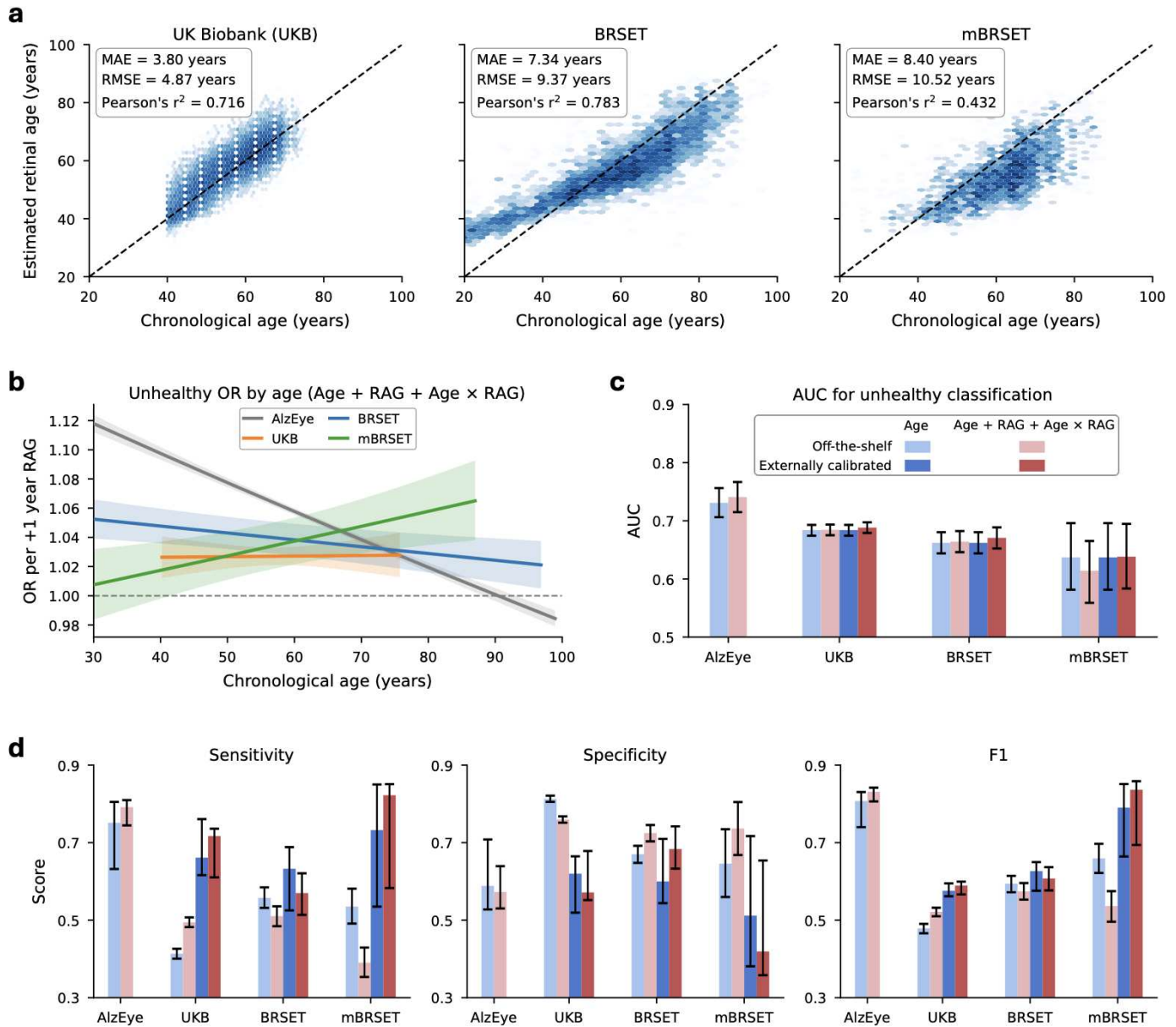


Fig. 5. External validation across UK Biobank, BRSET and mBRSET. **a.** External retinal age estimation performance. Hexbin density plots show estimated retinal age versus chronological age in each external test set. MAE, RMSE and Pearson's r^2 are annotated for each cohort. **b.** Age-specific odds ratio for unhealthy status per 1-year increase in RAG across cohorts. Curves were estimated from Age + RAG + Age $\times$ RAG logistic regression models fitted separately in each cohort and plotted across the corresponding chronological-age range. Shaded bands denote 95% CIs from patient-level cluster bootstrap. **c-d.** External disease classification performance for the unhealthy outcome. Two logistic regression models, Age only and Age + RAG + Age $\times$ RAG were evaluated in each cohort under two settings. Off-the-shelf: the AlzEye-fitted model was applied directly to the external test set, using the AlzEye internal Youden threshold for threshold-based metrics. Externally calibrated: the logistic regression model was refitted on the external training split and evaluated on the corresponding external test set, using a cohort-specific Youden threshold for threshold-based metrics. **c.** AUC. **d.** Sensitivity, specificity and F1 score at the corresponding operating threshold. Error bars denote 95% CIs from patient-level cluster bootstrap. Full results are reported in [Supplementary Tab. 10-12](#).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarytables.pdf](#)
- [supplementaryfigures.pdf](#)