npj Digital Medicine

https://doi.org/10.1038/s41746-026-02524-6

Article in Press

Time and person sensitive foundation model for disease prediction and risk stratification

Received: 7 October 2025

Accepted: 26 February 2026

Cite this article as: Wang, Z., Zhou, Y., Wu, Y. *et al.* Time and person sensitive foundation model for disease prediction and risk stratification. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02524-6

Zheyuan Wang, Yukun Zhou, Yilan Wu, Jocelyn Hui Lin Goh, Ke Zou, Zhouyu Guan, Yibing Chen, Gabriel Dawei Yang, Ping Zhang, Changchang Yin, An Ran Ran, Miao Li Chee, Can can Xue, Zhi da Soh, Samantha Yew, Danqi Fang, Xujia Liu, Benjamin Sommer Thinggaard, Jakob Grauslund, Haoxuan Li, Yixiao Jin, Jia Shu, Tingyao Li, Nan Jiang, Tingli Chen, Huating Li, Xiangning Wang, Qiang Wu, Charumathi Sabanayagam, Siegfried K. Wagner, Carol Y. Cheung, Ching-Yu Cheng, Bin Sheng, Tien Yin Wong, Pearse A. Keane & Yih-Chung Tham

Time and Person Sensitive Foundation Model for Disease Prediction and Risk Stratification

Zheyuan Wang[1,2†], Yukun Zhou[3,4,5†], Yilan Wu[6,3†], Jocelyn Hui Lin Goh[7,8†], Ke Zou[8], Zhouyu Guan[9], Yibing Chen[7], Gabriel Dawei Yang[7], Ping Zhang[10,11], Changchang Yin[10,11], An Ran Ran[12], Miao Li Chee[7], Cancan Xue[7], Zhida Soh[7], Samantha Yew[8], Danqi Fang[12], Xujia Liu[12], Benjamin Sommer Thinggaard[13], Jakob Grauslund[14], Haoxuan Li[15], Yixiao Jin[6], Jia Shu[1,2], Tingyao Li[1,2], Nan Jiang[1,2], Tingli Chen[16], Huating Li[9], Xiangning Wang[17], Qiang Wu[17], Charumathi Sabanayagam[7], Siegfried K Wagner[3,4], Carol Y. Cheung[11*], Ching-Yu Cheng[7,8*], Bin Sheng[1,2*], Tien Yin Wong[6,7,18*], Pearse A. Keane[3,4*], Yih-Chung Tham[7,8*]

[†]These authors contributed equally.

*These authors co-supervised this work: Carol Y. Cheung, Ching-Yu Cheng, Bin Sheng, Pearse A. Keane, Tien Yin Wong, Yih-Chung Tham

Corresponding authors: Bin Sheng(shengbin@sjtu.edu.cn), Pearse A. Keane (p.keane@ucl.ac.uk), Tien Yin Wong (wongtienyin@tsinghua.edu.cn), Yih-Chung Tham (thamyc@nus.edu.sg)

Affiliations

1. Department of Computer Science and Engineering, School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.
2. MOE Key Laboratory of AI, School of Electronic, Information, and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.
3. Institute of Ophthalmology, University College London, London, UK.
4. NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust, London, UK.
5. Hawkes Institute, University College London, London, UK.
6. Beijing Visual Science and Translational Eye Research Institute (BERI), Eye Center of Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China.
7. Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore.
8. Centre for Innovation and Precision Eye Health; and Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore.
9. Department of Endocrinology and Metabolism, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai Diabetes Institute, Shanghai Clinical Centre for Diabetes, Shanghai, China.
10. Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA.
11. Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA.

12. Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong Special Administrative Region, China.
13. Department of Ophthalmology, Odense University Hospital, Denmark
14. Department of Regional Health Research, University of Southern Denmark, Odense, Denmark.
15. Shanghai University of Sport, Shanghai, China.
16. Department of Ophthalmology, Shanghai Health and Medical Centre, Wuxi, Jiangsu, China.
17. Department of Ophthalmology, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China.
18. Beijing Visual Science and Translational Eye Research Institute, Beijing Tsinghua Changgung Hospital, Beijing, China.

ABSTRACT

Foundation models (FMs) enable generalizable medical AI, but existing retinal FMs perform best on cross-sectional classification and detection and are less effective for predicting disease incidence and progression. We present RETFound Plus, a CFP-based FM trained with temporal modeling on 1,304,292 fundus photographs from 304,345 participants across multiple visits to learn progression-aware representations. Compared with RETFound, RETFound Plus improved calibration and 5-year risk prediction across systemic and ocular diseases, with larger gains for systemic outcomes (stroke, myocardial infarction, diabetes and hypertension; +4–10% c-index) than ocular outcomes (diabetic retinopathy and glaucoma; +3–7% c-index), and improved risk stratification for systemic diseases (1.2–2.1-fold higher hazard-ratio trend). Results were consistent across external multi-regional, multi-ethnic datasets from the UK, US, Singapore, Hong Kong and Denmark. Clinical trial number: not applicable.

INTRODUCTION

Foundation models (FMs) represent a paradigm shift in artificial intelligence (AI)[1–3], demonstrating remarkable adaptability across diverse tasks through self-supervised learning on large-scale datasets. In medicine, these models have shown promise in ophthalmology[4], radiology[5], and pathology[6–8], achieving state-of-the-art performance in cross-sectional applications such as lesion detection and disease classification. However, their potential for predicting longitudinal outcomes, including disease incidence and progression, as well as risk stratification - remains largely underexplored[9].

The challenge of accurate temporal prediction in medicine stems from several unique factors. Unlike forecasting in other domains such as finance or climate science[10–12], medical prediction must account for complex, patient-specific biological processes that evolve non-linearly over time, as well as inter-individual variability in disease trajectory. These challenges are compounded by the frequent scarcity of high-quality longitudinal data and the common occurrence of incomplete medical records[13]. Whether existing FM, trained on cross-sectional data, can sufficiently perform temporal longitudinal prediction tasks well is unclear.

Color fundus photography (CFP) has been used as an exemplar in ophthalmology and medicine for real-world implementation of AI in medicine[14,15], not only for common eye diseases, but also as a tool for systemic health (a field termed "oculomics")[16,17]. CFP has advantages of being widely accessible even in community and low-resource settings, is non-invasive in nature, and provides rich *in vivo* information on the microvascular and neurological system. Multiple studies have demonstrated CFP's ability to predict onset and progression of systemic conditions including cardiovascular disease[18] and related risk factors including diabetes and hypertension[19], as well as predicting the prognosis of eye diseases[20]. Despite this potential, existing approaches for disease trajectory prediction remain limited by either reliance on non-imaging metadata alone[21] or narrow disease-specific focus[22].

To address these limitations, we introduce RETFound Plus, a novel FM that incorporates temporal encoding during pretraining to specifically enable longitudinal prediction. Our model captures patient-specific morphological changes in retina over time by pretraining on longitudinal sequences of retinal images, enabling a deeper understanding of disease progression. (**Figure 1**). This approach represents a potential significant advancement in disease prediction and risk

stratification over the current FMs, offering a unique opportunity to optimize personalized disease trajectory prediction.

To rigorously evaluate RETFound Plus, we benchmarked it against RETFound, and a Masked Autoencoder (MAE)[23] pretrained on ImageNet, and a series of state-of-the-art models, demonstrating superior performance of RETFound Plus in predicting incident ocular and systemic diseases over the current RETFound trained on cross-sectional data. Furthermore, using longitudinal clinical follow-up data from diverse, multi-ethnic, multinational cohorts, we demonstrated the longitudinal predictive performance of RETFound Plus for various systemic and eye diseases, as well as its risk stratification capabilities for systemic diseases. Finally, we demonstrated that incorporating temporal encoding during pretraining enhances RETFound Plus's model label efficiency, improving performance across varying data volumes, longitudinal durations, and image regions through ablation studies. Our result demonstrated that compared to prior FMs, RETFound Plus not only achieves superior longitudinal prediction performance and better risk-stratification performance, but it also exhibits greater generalizability, greater label efficiency to data sparsity, temporal availability, and masked imaging inputs.

By providing accurate prediction of disease onset, progression, and risk stratification, RETFound Plus could improve the use of AI as a clinician decision support tool for individualized disease management. Importantly, given the global scarcity of large-scale, high-quality longitudinal datasets, RETFound Plus can serve as a generalizable FM to be adapted across diverse populations and clinical settings. This has the potential to democratize access to disease prediction models, and advance AI-enabled precision medicine globally.

RESULTS

Prediction of systemic disease incidence and progression

To finetune and validate RETFound Plus for the prediction of systemic diseases, including diabetes, hypertension, CKD, MI, and stroke, we leveraged the longitudinal Diabetes Progression Study **(**DRPS[22]**)** cohort. This cohort includes individuals undergoing annual health screenings at Huadong Sanatorium and Shanghai Sixth People's Hospital. We independently curated progression data for each specific task from an original dataset of **19,100 participants**. We applied a unified data curation pipeline (**Supplementary Figure 5**) to construct task-specific datasets: (i) after image quality control, at-risk cohorts were defined by excluding individuals with pre-existing diseases at baseline; (ii) follow-up visits were linked longitudinally, with events and censoring defined by task-specific diagnostic criteria (**Supplementary Table 2**); (iii) final datasets were randomly sampled. Since each task was independently sampled, individual participants could appear across multiple datasets. Following inclusion and exclusion criteria, 11,938 participants were retained for model fine-tuning. Cohort sizes for each task are summarized in **Supplementary Table 1**.

For the internal validation, RETFound Plus showed about 3-4% higher 5-year c-index over RETFound on predicting diabetes (0.683 [95% CI, 0.667-0.699] vs 0.641 [95% CI, 0.622-0.661], P<0.001) and MI (0.944 [95% CI, 0.924-0.963 vs 0.919 [95% CI, 0.894-0.943], P<0.001) (**Figure 2**). The improvement margins of RETFound Plus were up to ~ 8% in predicting stroke (0.812 [95% 0.799-0.825] vs 0.738 [95% 0.722-0.753], P<0.001) and ~10% in predicting hypertension (0.753 [95% 0.744-0.763] vs 0.649 [95% 0.637-0.661], P<0.001). For CKD incidence prediction, RETFound Plus (0.693, 95% CI, 0.684-0.703) achieved a higher c-index than MAE (0.649, 95%CI 0.639-0.660, *P*<0.001) but was comparable with RETFound (0.691, 95%CI 0.681-0.702, *P*=0.223). In addition to disease incidence, RETFound Plus (0.633 [95% CI, 0.615-0.691]) also achieved the higher 5-year c-index for predicting the progression from pre-diabetes to diabetes when compared to RETFound (0.609 [95% CI, 0.593-0.626], *P*<0.001) (**Supplementary Table 3**). We also benchmarked RETFound Plus against a range of state-of-the-art (SOTA) models, including vision-language foundation models (CLIP[24], SigLIP[25], and MedCLIP[26]), a retinal multimodal foundation model (VisionFM[27]), and classical architectures (ResNet[28]-101, ConvNeXt[29]-Large, ViT[30]-Large), and also a deep learning survival model (DeepHit[31]). These results are shown in detail in **Supplementary Figure 2, Supplementary Table 12 and Supplementary Table 14.**

We further evaluated RETFound Plus's performance on multi-ethnic external datasets from various countries including the Chinese University of Hong Kong (CUHK) dataset from Hong Kong SAR (China), the Odense University Hospital (OUH) dataset from Denmark, the Age-Related Eye Disease Study (AREDS) from the US, the UK Biobank (UKB) and Moorfields Eye Hospital (MEH)-AlzEye from the UK, and the Singapore Epidemiology of Eye Diseases (SEED) dataset from Singapore. The external validation results showed the same trends as the internal results, with RETFound Plus showing significantly higher c-indexes for all tasks in all datasets (all P<=0.014) **(Supplementary Table 4),** and especially better performance on predicting the 5-year incidence of stroke (up to 0.127 higher c-index when comparing with RETFound), MI (up to 0.109 higher c-index when comparing with RETFound), and hypertension (up to 0.091 higher c-index when comparing with RETFound). These results collectively demonstrate the enhanced performance of RETFound Plus in predicting longitudinal disease incidence and progression outcomes.

Risk stratification for systemic diseases

In addition to disease prediction, we also analyzed the performance difference between RETFound Plus and its prior model RETFound from the aspect of risk stratification. Kaplan-Meier curves in the internal datasets for different systemic diseases were given in **Figure 3** and **Supplementary Figure 1**. Risk groups (Low, Medium, High) were defined based on the predicted risk scores from each model.

Although both RETFound and RETFound Plus showed significant divergence between the survival curves of the low-, medium-, and high- risk groups (with log-rank test P<0.001 in most tasks), the high-risk curve for RETFound Plus showed significantly steeper decline particularly around years 3-5 (**Figure 3**), indicating enhanced sensitivity to long-term outcomes.

Internal validation across different systemic diseases showed the significantly better risk-stratification ability of RETFound Plus (**Supplementary Table 5**), with about 1.05-2.10 folds higher trend hazard ratios (HRs) across all diseases when compared with RETFound. This can be attributed to both a lower incidence per person-year in the RETFound Plus low-risk group (0.72 for RETFound Plus, vs 2.28 for RETFound for stroke), and higher incidences per person-year in the RETFound Plus high-risk groups (19.47 for RETFound Plus, vs 15.19 for RETFound for stroke).

Similar results can be found in external validation datasets both for major cardiovascular diseases and other chronic diseases (**Supplementary Table 6**). RETFound Plus showed significantly better risk-stratification performance for stroke, achieving about 1.10-2.04 folds of trend HR when comparing with RETFound in the three external centres. For MI, the trend HRs of RETFound Plus were 1.26-1.54 folds higher than those of RETFound. For chronic diseases (including CKD, hypertension and diabetes), RETFound Plus also significantly better risk-stratification ability, with about 1.07-1.97 higher trend HRs higher than those of RETFound.

To conclude, RETFound Plus outperforms prior models in risk stratification, providing better separation between risk groups, and more precise incidence rate differentiation.

Prediction of ocular disease incidence and progression

In addition to systemic diseases, we also validated the model in predicting disease incidence and progression for ocular diseases.

In the internal validation (**Figure 4**), RETFound Plus generally achieved the best performance in most disease incidence prediction tasks. For instance, in DR incidence prediction and DME (diabetic macular edema) incidence prediction, RETFound Plus achieved a higher c-index than RETFound and MAE (all P<0.001). A similar trend was also found in disease progression prediction tasks including DME progression (from non-centre to centre-involved DME [ciDME]), and any DR progression (all P<0.001, **Supplementary Table 7**).

For the prediction of 5-year incidence of eye diseases across external tests, RETFound Plus achieved c-indexes between 0.727–0.799, significantly outperforming RETFound (c-indexes between 0.646–0.717, all P<=0.006, **Supplementary Table 8)**. Similarly, for the prediction of 5-year ocular disease progression prediction, RETFound Plus achieved c-indexes of 0.656-0.684 for any DR progression, 0.747–0.869 for DR progression to VTDR, and 0.695–0.746 for DME progression. It significantly outperformed both RETFound and MAE in most external validation datasets **(Supplementary Table 8)**.

Efficiency and robustness analysis

We quantified efficiency along two axes, (1) data efficiency (**Figure 5-A**; varying training set size) and temporal-information efficiency (**Figure 5-B**; varying the proportion of available time-

encoded information), and assessed robustness by analysing reasoning over disrupted spatial dependencies (**Figure 5-C**).

Label efficiency for data volume measures the amount of training data and annotations required to achieve a given performance on downstream tasks, reflecting the annotation workload for medical experts. Specifically, RETFound Plus and RETFound were trained on subsets of the training dataset (10%, 20%, 50%, 90%, and 100%) and evaluated on the whole test set. The results showed that RETFound Plus consistently outperformed RETFound across all tasks with less training data (**Figure 5-A**). Notably, for stroke incidence, RETFound Plus achieved a comparable c-index with only 20% of training data (0.733, 95%CI 0.711-0.756) to RETFound with 100% proportion of training data (0.725, 95%CI 0.694-0.756) (**Supplementary Table 9**). Similar result was shown in DME and glaucoma incidence prediction where RETFound Plus with 20% of training data achieved similar c-index with RETFound with 100% training data (**Supplementary Table 11**). While for hypertension and diabetes incidence **(Supplementary Table 9)**, as well as for DR incidence prediction (**Supplementary Table 11**), RETFound Plus with only 50% of training data achieved a comparable c-index with RETFound with 100% of training data. These results demonstrated its advantage of leveraging less training data for more accurate longitudinal prediction.

Collecting data from longitudinal cohorts is always expensive and highly time and labour-consuming, contributing to the scarcity of longitudinal cohorts with long follow-up periods. To demonstrate that RETFound Plus had good label efficiency when handling longitudinal datasets of varying durations, we analyzed label efficiency for the proportion of available time information by finetuning the models on training datasets with varying maximum follow-up durations. Generally, both RETFound and RETFound Plus showed continuously increasing c-indexes as the cut-off duration elongated, which demonstrated the importance of training datasets with sufficient longitudinal follow-up periods to the performance of models. Across different downstream tasks, RETFound Plus consistently outperformed RETFound across different follow-up periods, as shown in **Figure 5-B**. RETFound Plus trained with 2-year follow-up dataset showed c-index of 0.652 (95%CI 0.640-0.664) in predicting the 5-year incidence of hypertension, comparable with RETFound trained with 5-year follow-up dataset (0.644, 95%CI 0.634-0.655) (**Supplementary Table 9**). The advantage of RETFound Plus in longitudinal label efficiency was also significant for ocular diseases, using 2-year follow-up dataset to achieve higher c-index when comparing with

RETFound for 5-year incidence prediction of both DME and glaucoma (**Supplementary Table 10**). When focusing on the performance platform, RETFound Plus trained with 2-year follow-up dataset showed c-index of 0.762 (95%CI 0.738-0.785) in predicting the 5-year incidence of stroke, comparable with it trained with 5-year follow-up 0.785 (95%CI 0.768-0.802), and also comparable with RETFound trained with 5-year follow-up dataset (0.728, 95%CI 0.697-0.759) (**Supplementary Table 9**).

Finally, we evaluated the stability of attention modelling using a controlled occlusion paradigm that randomly masks multiple 16×16 patches (matching the Transformer patch size) at varying ratios, thereby probing the models' ability to capture feature dependencies under spatial perturbations. Performance declined across tasks as the occlusion ratio increased; nevertheless, RETFound Plus consistently outperformed RETFound **(Figure 5)**, indicating greater robustness to spatial disruption. Under severe occlusion (70% of patches masked), RETFound Plus achieved relative gains of approximately 4–7% for stroke, myocardial infarction, pre-diabetes and diabetes prediction, and about 2.5% and 1% for hypertension and chronic kidney disease, respectively (**Supplementary Table 11**). These findings suggest that RETFound Plus more effectively reasons over disrupted spatial dependencies. Its multi-head self-attention reweights attention towards unoccluded regions, enabling the extraction of informative signals from residual semantic patches even under substantial information loss.

DISCUSSION

In clinical practice, physicians are required to provide diagnostic and management plans to the patient, offering options of differential diagnosis, initiating further tests or treatment options, and providing a reasonable forecast of the patient's future disease state and prognosis. However, current FMs, including those in ophthalmology[4], pathology[6], and oncology[8], are mainly designed for the initial tasks of lesion detection and disease diagnosis, and fail to adequately address the other key clinical needs of predicting disease onset, progression, and disease prognosis. To bridge these gaps, we introduce RETFound Plus, a longitudinal FM that captures individual-specific retinal changes and is sensitive to temporal dynamics. We first benchmarked RETFound Plus with RETFound, by evaluating its predictive performance for ocular and systemic diseases, tested internally and across multiple external datasets. We demonstrated that RETFound Plus outperformed RETFound with notable performance gains for predicting systemic disease (e.g., improvement in 4% to 10% in predicting the 5-year incidence and progression of systemic diseases). In label efficiency evaluation, we demonstrated that RETFound Plus maintained strong performance even with 20% less data, 1-year shorter follow-up, and under condition of increased retinal image occlusion area (20%). RETFound Plus also showed superior performance to RETFound in predicting ocular diseases, though the improvement margin was more modest compared to systemic disease outcomes. Taken together, these findings provide unique insights into the value of temporally informed representation learning in the development of FMs for longitudinal risk prediction, and highlight the potential of future FMs to incorporate temporal information to advance personalized and preventive medicine.

While traditional AI models have shown reasonable performance in prediction of specific disease incidence and progression (e.g., DR[22], stroke[32], neurodegenerative diseases[33], and tumors[34]), these models are not only disease- but also task-specific. However, it is unrealistic to develop new models for each disease, condition and task, as large longitudinal cohort data to train such traditional AI models are scarce. A prior study presented an unsupervised technique that did not require specific domain knowledge and could accommodate to different fields[18]; however, this model was limited by utilizing data from only electronic health records. To address this gap and provide a foundational self-supervised model applicable to a wider range of disease and tasks, our RETFound Plus framework expands upon the previous cross-sectional RETFound model by learning the trajectories of follow-up data. By incorporating temporal data at pretraining stage,

RETFound Plus can account for the dynamic nature of disease progression over time, thereby enhancing its predictive accuracy as we have shown. RETFound Plus also utilizes a combination of reconstruction loss and contrastive loss. This approach enables RETFound Plus not only to maintain image reconstruction capabilities but also to learn the temporal correlations within individual instances of data. As a result, RETFound Plus could achieve more robust and accurate predictions than traditional AI models or current FMs using cross-sectional data (i.e., RETFound).

Diseases often manifest and evolve in a non-linear and patient-specific manner, with various influencing factors that change over time[35]. Precision medicine is crucial for personalized cardiovascular care[36]. However, traditional risk scores may not be precise for individual-level predictions[37-39]. By leveraging contrastive loss to learn the relationships between CFP of the same eye captured at different time points during the pretraining phase, RETFound Plus is better aligned with the real-world dynamics of disease progression. We show this capability enhances its ability to predict future clinical outcomes based on both historical and current data. The time-aware feature representations through this approach not only improve predictive performance but also enable more personalized risk assessments, which are important in clinical settings for effective management of chronic diseases progression over time. For example, RETFound Plus improved the performance of prior FM in predicting the 5-year disease incidence by higher c-indexes of about 4-5% for diabetes and MI, while 7-10% for stroke and hypertension. Although the predictive performance of all models for DM and HTN incidence was low, this may be attributed to underreporting in DM and HTN cases, with approximately half remaining undetected[40]. Vision impairment and blindness represent significant public health challenges, imposing substantial burdens on both individual quality of life and healthcare systems[41]. Conditions such as AMD, glaucoma and DME significantly contribute to the prevalence of vision impairment, with projections indicating a rise in prevalence owing to global population aging[42]. RETFound Plus improved the performance of prior FM in predicting the 5-year disease incidence by higher c-indexes of about 5% for AMD, 6% for glaucoma and 7% for DME. In this context, RETFound Plus has the potential to impact major eye diseases by facilitating personalized screening and follow-up intervals, personalized risk stratification and treatment strategies[43].

Accurate disease risk stratification is critical for enabling precision medicine, as it identifies high-risk individuals who may benefit from early interventions while avoiding overtreatment of low-risk populations. Our results showed that RETFound Plus consistently demonstrates stronger

risk discrimination across all outcomes, with higher HRs for high-risk groups compared to RETFound. For MI and stroke, RETFound Plus shows particularly large improvements, with HRs >10 for high-risk patients, compared to RETFound's HRs of ~4–8. In addition to HRs, the incidence rates also show sharper contrasts between low- and high-risk groups in RETFound Plus, suggesting better model calibration. Thus, we think the strong performance of RETFound Plus in risk stratification has great clinical promise in personalized risk assessment and preventive medicine.

We systematically evaluated the impact of training data quantity, follow-up duration, and image quality on the performance of RETFound Plus to address the difficulty of longitudinal clinical data scarcity. While RETFound Plus exhibited a performance decline under these data limitations, it still demonstrated a consistently higher c-index compared to the prior foundation model. For instance, RETFound Plus trained with 2-year follow-up dataset showed c-index of 0.652 (95%CI 0.640-0.664) in predicting the 5-year incidence of hypertension, comparable with RETFound trained with 5-year follow-up dataset (0.644, 95%CI 0.634-0.655). These findings suggest that RETFound Plus exhibits greater label efficiency in real-world clinical settings, highlighting its potential for enhanced translatability to future clinical applications.

To rigorously validate the individual contributions of our pre-training strategy, we conducted comprehensive ablation studies dissecting the impact of the dual-loss formulation and the explicit modeling of temporal disease evolution. We compared a baseline model trained only with the Masked Image Modeling (MIM) loss, a "Dual-Loss without Temporal Modeling" variant that adopts the teacher–student architecture but treats longitudinal images as independent cross-sectional samples (thus removing temporal correspondence), and our full RETFound Plus framework, which combines the dual-loss mechanism with longitudinal temporal modeling to learn from the subtle progression of the same eye over time. As presented in **Supplementary Figure 4** and **Supplementary Table 15**, the Dual-Loss without Temporal Modeling variant consistently improved performance over the single-loss baseline across several tasks, including stroke, MI, and HTN incidence prediction, suggesting that the teacher–student dual-loss formulation intrinsically enhances feature representation even in the absence of temporal pairing. Nevertheless, the complete RETFound Plus model achieved significantly superior performance across all downstream tasks compared with both the baseline and the non-temporal variant ($P < 0.05$), indicating that the main advantage of our approach arises not merely from architectural

refinements but from the effective exploitation of longitudinal information. By explicitly guiding the student network to predict future states from past observations, RETFound Plus captures critical disease progression patterns that are largely inaccessible to purely cross-sectional self-supervised approaches.

We conducted subgroup analyses for age, sex, and imaging device on HTN incidence task (**Supplementary Figure 3** and **Supplementary Table 5**). Notably, the model performed better on images from the TRC NW300 than the TRC NW400, and achieved higher C-indices in females than in males, indicating meaningful performance gaps across both acquisition hardware and demographic subgroups. These findings underscore that diverse, well-balanced training data across devices and populations are essential for improving model fairness and ensuring robust, equitable deployment of RETFound Plus.

For transparency, we included longitudinal attention trajectories and visit-level attention visualizations in **Supplementary Figures 6**, **7**, and **8**. Across tasks and examples, attention maps generally highlighted clinically plausible retinal regions (e.g., optic disc, macula, and vasculature), and predicted survival curves showed time-localized changes that were qualitatively compatible with subsequent event occurrence. Nevertheless, in our current analyses, these fine-grained temporal attention trajectories did not provide additional consistently reliable or clinically actionable information beyond a qualitative sanity check, and the relationship between dynamic attention patterns and the model's progression predictions remained difficult to interpret mechanistically. We therefore presented these figures as supportive visual documentation of model behavior rather than as evidence of a validated biomarker-level explanation, and further work was needed to establish whether and how attention dynamics could be translated into clinically meaningful temporal interpretability.

Our study has several limitations. First, within the RETFound Plus framework, the teacher encoder requires full-resolution images as input, and the student encoder operates with a lower masking ratio compared to RETFound. These design choices significantly increase the computational resources needed for training, which may not be accessible to all research teams or institutions. This constraint could limit the broader adoption and further development of the proposed method. To that end, future study is warranted to explore more computationally efficient training strategies or model architectures that can maintain performance while reducing resource demands. Second, the pretraining data used in this study primarily consists of longitudinal CFPs

collected from specific populations or clinical environments. This could restrict the model's generalizability to populations with different demographic, genetic, or environmental characteristics. To address this limitation, future studies should incorporate more diverse and representative datasets, which would help mitigate potential biases and enhance the model's robustness. Third, as in other studies, the performance of RETFound Plus in external validation was not as strong as in the internal dataset. This may be explained by the different follow-up interval between different cohorts and datasets. The dataset we used to pretrain and finetune RETFound Plus model was from a community-based cohort in a seven-year period with one-year interval between each follow-up. However, the external validation datasets include population-based cohort with much longer interval (such as SEED and AREDS), as well as hospital-derived cohort with irregular interval (such as MEH and OUH). These different characteristics including follow-up time across datasets, might explain the performance drop in external validations. For example, population-based cohorts with longer follow-up interval may have 'delayed record of the time of 'actual' disease incidence and progression. On the other hand, while hospital or clinical cohort may have more accurate time-to-event record, the follow-up intervals are irregular between patients. Future study with foundation model finetuned with more diverse data might be helpful to answer this question. While this study focuses on retrospective longitudinal validation, future prospective studies are needed to evaluate how clinicians interpret and act upon predicted long-term risk estimates within a human-in-the-decision-loop framework, and whether such integration improves downstream clinical outcomes.

In summary, we developed a novel FM, RETFound Plus, which was designed to encode longitudinal information in the pre-training phase of model development and to learn the correlations between CFPs acquired across different time points. We showed RETFound Plus demonstrates an ability to more accurately predict disease risk and progression comparing with prior FMs, including RETFound, offering the potential for personalized risk stratification and disease progression monitoring based on baseline retinal imaging alone. Importantly, given the global scarcity of large-scale, high-quality longitudinal datasets, RETFound Plus can serve as a generalizable FM to be fine-tuned and adapted across diverse populations and clinical settings. RETFound Plus thus represents a significant advancement in longitudinal disease prediction and risk stratification and holds promise for enhancing clinical decision support, personalized care and population health strategies globally.

METHODS

Data acquisition and disease definition

RETFound Plus was pretrained on a longitudinal dataset consisting of Shanghai Diabetes Prevention Program[22], Nicheng Diabetes Screening Project[35] and Shanghai Integration Model[44] collected between January 2015 and January 2024, including 1,304,292 CFPs from 304,345 participants. The mean age for patients was 58.07 ±14.02 years old, and 35.3% of the participants were female. Diabetes Progression Study (DRPS) cohort was used to finetune the RETFound Plus model. The DRPS[22] is a longitudinal cohort of participants who underwent annual health examination in Huadong Sanatorium and Shanghai Sixth People's Hospital in a 5-year period. We included 11,938 participants in this study for finetuning. The diagnosis standard of each disease in DRPS was listed in detail in **Supplementary Table 2**.

External validation included multi-country datasets. Macular-centered retinal photographs were captured for each eye at baseline and follow-up visits were used in all external datasets. The CUHK-STDR cohort[45] was a prospective observational study involving 337 patients with diabetes. Participants were recruited from CUHK Eye Centre in Hong Kong between July 2015 and November 2016 and had been consecutively followed up for at least 5 years. The SEED cohort[46] is a multiethnic longitudinal population-based study including Singaporean adults of Malay, Indian and Chinese descent. In total, 6,188 individuals with diabetes from the SEED cohort with 6-year follow-ups were included for external validation. The OUH cohort[47] was a retrospective hospital-derived cohort involving 535 patients at Odense University Hospital, Odense, Denmark between 2015 and 2022. The AREDS cohort[48] was a randomized, placebo-controlled clinical trial based at 11 retinal specialty clinics in the United States that enrolled 4757 participants from November 1992 to November 2005. Eligible participants for the AREDS were aged 55 to 80 years and could have no AMD in either eye, early or intermediate AMD in one or both eyes, or late AMD in one eye only. Participants underwent annual visits and at visits every 6 months if progression to late AMD was suspected. The MEH-AlzEye[4] is a retrospective cohort study linking ophthalmic data of 353,157 patients, who attended Moorfields Eye Hospital between 2008 and 2018, with systemic health data from hospital admissions across the whole of England. Systemic health data are derived from Hospital Episode Statistics (HES) data relating to admitted patient care, with a focus on cardiovascular disease and all-cause dementia. Diagnostic codes in HES admitted patient care are

reported according to the tenth revision of the ICD (International Statistical Classification of Diseases)[4].

Disease incidence was defined as patients without disease at baseline who were diagnosed with disease at follow-up. For DR progression, we included three different DR progression outcomes – 1) DR any progression, defined as patients with any increase in DR severity at follow-up; 2) DR progression to referable DR, defined as patients without DR or referable DR who progressed to referable DR at follow-up; and 3) DR progression to vision-threatening DR (VTDR), defined as patients without DR or VTDR who progressed to VTDR at follow-up. DME progression was defined as patients with no DME or non-center-involved DME progressing to have center-involved DME at follow up. The progression of AMD was defined as eyes with non-late-AMD (early or intermediate AMD; AREDS category 2-3) at baseline but being graded as having late AMD (AREDS category 4) during a follow-up period. Detailed definition of tasks was shown in **Supplementary Table 1**, and disease definitions in each external validation cohort were shown in **Supplementary Table 2.**

Study design and data curation for self-supervised learning

We developed a longitudinal self-supervised learning framework, named RETFound Plus, based on fundus photography to capture temporal disease progression patterns. To achieve this, we curated a comprehensive longitudinal dataset, where color fundus photographs were systematically organized by individual participants, with each eye serving as an independent analytical unit across multiple follow-up visits. For each participant, we maintained comprehensive temporal records of all available fundus images spanning multiple visits, establishing participant-specific temporal caches indexed by eye laterality and visit timestamps. This organizational structure enabled efficient temporal pairing during self-supervised training while preserving anatomical consistency and temporal ordering. All fundus images were preprocessed using circular cropping to remove non-retinal background. Image quality was assessed by the EyeQ[49], and only images graded as “good” or “usable” were retained. For each visit, the image with the highest quality score was selected if multiple images were available.

Temporal self-supervised learning framework

Our temporal self-supervised framework adopts a dual-encoder design[50] to capture intra-ocular temporal dynamics, leveraging asymmetric temporal inputs and hierarchical supervision (as shown in **Figure** 1). Specifically, the student encoder processes masked images from time point t_1, while the teacher encoder receives unmasked images from both t_1 and t_2, so that the teacher can provide both patch-level targets at t_1 and global temporal targets at t_2, enforcing consistency between current and future representations. Both encoders utilize Vision Transformer (ViT-Large/16) architectures. The teacher network parameters θ_t are updated via exponential moving average of the student parameters θ_s: $\theta_t \leftarrow \tau\theta_s + (1-\tau)\theta_t$, where $\tau = 0.996$. This momentum teacher design, similar to that used in BYOL[51], DINO[52], means that the teacher is not an externally supervised model but a slowly updated, temporally smoothed version of the student, which serves as a stable target network rather than a conventional teacher that is trained independently. This asymmetric update stabilizes training and mitigates representation collapse while still allowing information to flow from the student to the teacher in a controlled manner. This asymmetric design enables stable learning while preserving temporal consistency.

Our framework implements a dual-level supervision mechanism that captures both local reconstruction and global disease progression patterns. The first supervision is patch-level reconstruction. Here, the student encoder receives t_1 images with randomly applied patch-wise masking (masking ratio $r \in [0.1,0.5]$), while the teacher encoder processes the corresponding unmasked t_1 images. Importantly, the reconstruction target at time t_1 is the same image at t_1 (not an image from a different time point t_2), which avoids relying on explicit cross-time image registration and circumvents potential spatial misalignment or displacement between t_1 and t_2 caused by changes in imaging conditions or patient positioning. This configuration enables the model to learn local dependencies by reconstructing masked regions using supervision from the teacher's patch-level representations. The temporal masked image modeling loss is defined as:

$$\mathcal{L}_{\text{MIM}} = -\sum_{i=1}^{N} m_i \cdot P_{\text{teacher}}^{\text{patch}}(x_{t_1,i})^T \log P_{\text{student}}^{\text{patch}}(\hat{x}_{t_1,i}) \quad (1)$$

where m_i indicates masked patches and P^{patch} represents patch-level probability distributions. By conditioning on unmasked teacher features at t_1, this objective focuses on learning robust local retinal structures and intra-image dependencies without requiring inter-temporal spatial alignment.

Secondly, to capture long-term disease progression patterns, we implement global temporal supervision using [CLS] tokens. The teacher encoder processes t_2 images to generate global representations that supervise the student's [CLS] token derived from masked t_1 images:

$$\mathcal{L}_{\text{GTC}} = -P_{\text{teacher}}^{[\text{CLS}]}(x_{t_2})^T \log P_{\text{student}}^{[\text{CLS}]}(\hat{x}_{t_1}) \quad (2)$$

Here, the goal is to encourage the global representation of the current visit (t_1) to approximate the global representation of a future visit (t_2) for the same eye. Because this loss operates on high-level [CLS] embeddings rather than on pixels or patches, it is inherently more tolerant to small shifts, scale changes, or field-of-view differences between t_1 and t_2, and thus does not require explicit image registration across visits. This design enables the model to learn predictive representations that anticipate future disease states from current observations. In other words, the student is explicitly guided to encode at t_1 those features that are most informative about the later disease status observed at t_2, which is crucial for modeling temporal disease evolution.

Training configuration

We randomly selected two images from the same eye of identical participants captured at different time points. The combined objective function $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MIM}} + \mathcal{L}_{\text{GTC}}$ was optimized using AdamW[53] with a base learning rate of 1.5×10^{-4}, scaled proportionally to batch size. Training employed a linear warm-up schedule over 10 epochs, followed by cosine annealing across 200 epochs. Pretraining was conducted on six NVIDIA A800 SXM4 80GB GPUs with a per-GPU batch size of 60 (total batch size 360).

Adaptation to downstream tasks

For the downstream tasks of disease risk and progression prediction, RETFound Plus was fine-tuned using a pseudo-survival model[54]. Each data point was represented as a triple $(\boldsymbol{x}, t, e)$, where $\boldsymbol{x}$ is the CFP at baseline, t is either the time of event occurrence or the time of censoring, and e is an event indicator (1 if the event occurred before censoring, 0 otherwise). For image *i*, the model estimated a survival probability function $S_i(t)$, representing the probability that the event time exceeds t. This survival function was modeled using a log-logistic distribution with patient-specific location μ_i and scale σ_i parameters. The parameters were derived from the high-

dimensional representations output by the student encoder, which were subsequently processed by a fully connected layer to predict the patient-specific parameters μ_i and σ_i.

The fine-tuning process was optimized using a negative log-likelihood loss function, which was specifically designed to handle censored data. The survival function $S_i(t)$ was expressed as:

$$S_i(t;\mu_i,\sigma_i) = \frac{1}{1+exp\left(\frac{logt-\mu_i}{\sigma_i}\right)} \quad (3)$$

The corresponding log-likelihood function was minimized to estimate the parameters μ_i and σ_i. The loss function accounted for both censored and uncensored data by combining the likelihood of uncensored events and the survival probability of censored observations.

We performed fine-tuning for 50 epochs with a batch size of 16. Optimization was carried out using a base learning rate of 2×10^{-4}, a layer-wise learning rate decay of 0.65, and a weight decay of 0.05. To prevent overfitting, we employed a drop path rate of 0.2. We also implemented a robust data augmentation pipeline during training, which included random rotation, random resized cropping (scale 0.64–1.0 with bicubic interpolation), random horizontal flipping, and color jittering (brightness, contrast, and saturation of 0.15).

Evaluation and statistical analysis

Performance of the models was evaluated using two metrics. The first was Harrell's c-index, using the patient-specific μ_i parameters as risk scores ($exp(\mu_i)$ to assess risk discrimination ability. The second was the integrated Brier score, which measures the time-averaged mean squared error between true binary outcomes and predicted probabilities, evaluating both calibration and discrimination. Both metrics were adjusted for censoring by weighting with the inverse probability of censoring and were calculated up to a cutoff time τ (default $\tau = 10$ years, matching the dataset's maximum event time). To assess statistically significant differences (P<0.05) between RETFound Plus and the benchmark models for each task, bootstrap resampling was performed 1000 times, and the p-value was calculated using a two-sided t-test. The arithmetic mean of the concordance index and integrated brier score, across all external validation centers was computed for each year.

Ethics statement

This study was approved by the Ethics Committee of Shanghai Sixth People's Hospital (Approval No. 2019-087) and conducted in accordance with the Declaration of Helsinki. Only de-identified retrospective data were used for research, without the active involvement of patients. Informed consent was obtained from participants according to the requirement of each included cohort.

Data availability

The datasets used for pretraining, fine-tuning and validation in this study contain sensitive human participant information, including retinal images with potentially identifiable retinal vascular patterns, as well as linked clinical and follow-up records. To protect participant privacy and comply with the data governance requirements and ethics approvals of each contributing cohort and institution, these datasets are not publicly available. Access to the data may be granted for research purposes on reasonable request to the corresponding author(s), subject to approval by the relevant data custodians and/or institutional review boards, completion of applicable data use agreements, and any additional restrictions imposed by the originating cohorts (including for externally sourced datasets).

Code availability

The code used to train, fine-tune, and evaluate the model in this study is available at https://github.com/jaranwayne/RETFound-Plus.git. The environment was configured with the following dependencies: Python v3.8.2, Torch v.1.9.1, Torchvision v.0.10.1, Scikit-Image v.0.19.3, Scikit-Learn v.1.3.2, seaborn v.0.11.2, timm v.0.5.4, SciPy v1.10.1, opencv-python v.4.7.0.72, Pillow v9.5.0, setuptools v.59.4.0, Matplotlib v3.7.1, NumPy v1.24.2, Pandas v2.0.0, openpyxl v.3.1.2, and pycm v.4.0.

Acknowledgements

This study was supported by the National Key R&D Program of China (2022YFC2502800), the National Natural Science Foundation of China (82388101) and the Beijing Natural Science Foundation (IS23096) to T.Y.W.; the National Medical Research Council of Singapore (NMRC/MOH/HCSAINV21nov-0001) to Y.-C. T.; the National Natural Science Foundation of China (62272298), the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0509202 & 2023ZD0509201), the National Key Research and

Development Program of China (2022YFC2407000) to B.S.; Wellcome Award 318987/Z/24/Z to Y.Z.; and the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0509202 & 2023ZD0509201), the Clinical Special Program of Shanghai Municipal Health Commission (20224044) and the Three-Year Action Plan to Strengthen the Construction of the Public Health System in Shanghai (2023-2025 GWVI-11.1-28) to T.C.

Author Contributions

Y-C.T., B.S. and T.Y.W. conceptualized this work. Z.W. developed the algorithm. Z.W., Y.Z. conducted the experiments. Z.W., Y.W. analyzed the data, prepared the figures and tables, and drafted the manuscript. J.H.L.G., K.Z., Y.C., Z.G., Y.C., G.D.Y., P.Z., C.Y., A.R.R., M.L.C., C.X., Z.S., S.Y., D.F., X.L., B.T., J.G., H.L., Y.J., N.J., H.L., J.S., T.L., T.C., X.W., Q.W., C.S., S.K.W., C.Y.C., C-Y.C., and P.A.K. contribute with the validation datasets. All authors reviewed the manuscript.

Competing Interests

P. A. K. is a cofounder of Cascader Ltd. and has acted as a consultant for Retina Consultants of America, Roche, Boehringer-Ingleheim, and Bitfount and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Thea, Apellis, and Roche. He has received travel support from Bayer and Roche. He has attended advisory boards for Topcon, Bayer, Boehringer-Ingleheim, and Roche. T. Y. W. is a consultant for Abbvie Pte Ltd, Aldropika Therapeutics, Bayer, Boehringer-Ingelheim, Carl Zeiss, Genentech, Iveric Bio, Novartis, Opthea Limited, Plano, Quaerite Biopharm Research Ltd, Roche, Sanofi, Shanghai Henlius. He is an inventor, holds patents and is a co-founder of start-up companies EyRiS and Visre, which have interests in, and develop digital solutions for eye diseases. All potential conflicts of interests for consultancy, advisory boards and positions in the start-up companies, and financial remuneration, if any, are managed by institutional policies under SingHealth and Tsinghua University. The other authors declare no financial or non-financial competing interests.

Figure Legends

Figure 1. Illustration of the study design

A. Data curation in longitudinal cohorts. CFPs were systematically organized into longitudinal cohorts, with each eye serving as an independent unit of analysis. Each image was annotated with its corresponding follow-up time and disease label.
B. Self-supervised on longitudinal CFPs. To model intra-eye temporal progression, we developed a self-supervised learning framework that leverages image pairs from the same eye captured at different time points. The framework consists of a student network and a teacher network. The student network receives a partially masked image, while the teacher network processes the unmasked counterpart and is updated via an exponential moving average (EMA) of the student weights. The model is jointly optimized with two loss functions: a global consistency loss aligns [CLS] token representations to capture longitudinal semantic changes, and a masked image reconstruction loss encourages learning of local pathological features. This approach enables explicit modeling of retinal temporal dynamics and facilitates transfer to a wide range of clinical tasks.
C. Fine-tuning and external validation. The pretrained student encoder of RETFound Plus was fine-tuned on labeled datasets to enhance its performance for downstream clinical tasks. External validation was conducted across multiple countries and ethnic groups, covering a broad range of applications, including predicting the incidence and progression of ocular and systemic diseases. The event was defined as the first re-visit demonstrating any stage of disease progression, with time to event measured as the interval from baseline to this visit.
DM, diabetes mellitus; CKD, chronic kidney disease; MI, myocardial infarction; RFP, RETFound Plus. SDPP, Shanghai Diabetes Prevention Program.

Figure 2. Internal validation results for the c-index of 5-year incidence or progression of systemic diseases.

DM, diabetes mellitus; CKD, chronic kidney disease; MI, myocardial infarction; MAE, masked autoencoder.
The error bars show 95% CI and the bar centre represents the mean value. *** represents P<0.001 when compared with RETFound Plus.

Figure 3. Kaplan-Meier curve for cumulative survival probability of systemic diseases in internal validation results.

RETFound and RETFound Plus were used for risk stratification for different systemic diseases. Risk groups (Low, Medium, High) were defined based on the tertiles of predicted risk scores from each model. The x-axis represents the follow-up years, and the y-axis shows the survival probability of different systemic disease outcomes.
DM, diabetes mellitus; CKD, chronic kidney disease; MI, myocardial infarction

Figure 4. Internal validation results for the c-index of 5-year incidence or progression of ocular diseases.

AMD, age-related macular degeneration; DR, diabetic retinopathy; DME, diabetic macular edema; VTDR, vision-threatening diabetic retinopathy; MAE, masked autoencoder.
The error bars show 95% CI and the bar centre represents the mean value. *** represents P<0.001 when compared with RETFound Plus.

Figure 5. Ablation experiments for label efficiency

A. Label efficiency for data volume. This figure shows the 5-year C-index of the RETFound and RETFound Plus models, with different percentages of training data (ranging from 10% to 100%). The performance for each data size is represented by the solid lines, with confidence intervals shown as shaded regions. The x-axis represents the percentage of training data used, and the y-axis shows the C-index.
B. Label efficiency from available of time information. This figure shows the 5-year C-index of the RETFound and RETFound Plus models, across different follow-up period cutoffs in a survival analysis task. The models were trained using subsets of the training data, filtered by follow-up period less than 1, 2, 3, 4, and 5 years, respectively, while the test set remained unchanged. The x-axis represents the cutoff period (in years), and the y-axis shows the C-index. The performance is represented by the solid lines, with confidence intervals shown as shaded regions.
C. Robustness to spatial occlusions. This figure shows the 5-year C-index of the RETFound and RETFound Plus models, across different image masking ratio. In this setting, random masking is performed by randomly selecting a certain proportion of the ViT input patches and setting them to zero, thereby simulating different degrees of information loss. The performance is represented by the solid dots.
C-index, concordance index; DM, diabetes mellitus; CKD, chronic kidney disease; MI, myocardial infarction.

Reference

1. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
2. Azizi, S. *et al.* Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat. Biomed. Eng.* **7**, 756–779 (2023).
3. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* **6**, 1346–1352 (2022).
4. Zhou, Y. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
5. Wang, J. *et al.* Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat. Med.* https://doi.org/10.1038/s41591-024-03359-y (2024) doi:10.1038/s41591-024-03359-y.
6. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Med.* https://doi.org/10.1038/s41591-024-02857-3 (2024) doi:10.1038/s41591-024-02857-3.
7. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat. Med.* https://doi.org/10.1038/s41591-024-02856-4 (2024) doi:10.1038/s41591-024-02856-4.
8. Vorontsov, E. *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* https://doi.org/10.1038/s41591-024-03141-0 (2024) doi:10.1038/s41591-024-03141-0.
9. Han, T. *et al.* Image prediction of disease progression for osteoarthritis by style-based manifold extrapolation. *Nat. Mach. Intell.* **4**, 1029–1039 (2022).
10. Mukkavilli, S. K. *et al.* AI Foundation Models for Weather and Climate: Applications, Design, and Implementation. Preprint at http://arxiv.org/abs/2309.10808 (2023).
11. Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. ClimaX: A foundation model for weather and climate. Preprint at http://arxiv.org/abs/2301.10343 (2023).
12. Wang, X., Feng, M., Qiu, J., Gu, J. & Zhao, J. From News to Forecast: Integrating Event Analysis in LLM-Based Time Series Forecasting with Reflection. Preprint at http://arxiv.org/abs/2409.17515 (2024).
13. Wang, X., Sontag, D. & Wang, F. Unsupervised learning of disease progression models. in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* 85–94 (ACM, New York New York USA, 2014). doi:10.1145/2623330.2623754.
14. Ting, D. S. W. *et al.* Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* **318**, 2211 (2017).
15. Gunasekeran, D. V. *et al.* National Use of Artificial Intelligence for Eye Screening in Singapore. *NEJM AI* **1**, (2024).
16. Wagner, S. K. *et al.* Insights into Systemic Disease through Retinal Imaging-Based Oculomics. *Transl. Vis. Sci. Technol.* **9**, 6–6 (2020).
17. Wang, J. *et al.* Artificial intelligence-enhanced retinal imaging as a biomarker for systemic diseases. *Theranostics* **15**, 3223–3233 (2025).
18. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
19. Ong, J. *et al.* Development of oculomics artificial intelligence for cardiovascular risk factors: A case study in fundus oculomics for HbA1c assessment and clinically relevant considerations for clinicians. *Asia-Pac. J. Ophthalmol.* **13**, 100095 (2024).
20. Holste, G. *et al.* Harnessing the power of longitudinal medical imaging for eye disease prognosis using Transformer-based sequence modeling. *Npj Digit. Med.* **7**, 216 (2024).
21. Qiu, J. *et al.* Deep representation learning for clustering longitudinal survival data from electronic health records. *Nat. Commun.* **16**, 2534 (2025).
22. Dai, L. A deep learning system for predicting time to progression of diabetic retinopathy. *Nat. Med.*
23. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. Preprint at http://arxiv.org/abs/2111.06377 (2021).
24. Radford, A. *et al.* Learning Transferable Visual Models From Natural Language Supervision. in *Proceedings of the 38th International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
25. Zhai, X., Mustafa, B., Kolesnikov, A. & Beyer, L. Sigmoid Loss for Language Image Pre-Training. in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 11941–11952 (2023). doi:10.1109/ICCV51070.2023.01100.

26. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. *Proc. Conf. Empir. Methods Nat. Lang. Process. Conf. Empir. Methods Nat. Lang. Process.* **2022**, 3876–3887 (2022).
27. Qiu J, Wu J, Wei H, et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence[J]. *NEJM AI*, 2024, 1(12): AIoa2300221.
28. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016). doi:10.1109/CVPR.2016.90.
29. Liu, Z. *et al.* A ConvNet for the 2020s. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11966–11976 (2022). doi:10.1109/CVPR52688.2022.01167.
30. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. in (2020).
31. Lee, C., Zame, W., Yoon, J. & Schaar, M. van der. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proc. AAAI Conf. Artif. Intell.* **32**, (2018).
32. Allen, A. *et al.* A Digital Twins Machine Learning Model for Forecasting Disease Progression in Stroke Patients. *Appl. Sci.* **11**, 5576 (2021).
33. Young, A. L. *et al.* Data-driven modelling of neurodegenerative disease progression: thinking outside the black box. *Nat. Rev. Neurosci.* **25**, 111–130 (2024).
34. Chen, R. & Wang, H. Time-to-Event Endpoints in Imaging Biomarker Studies. *J. Magn. Reson. Imaging* jmri.29446 (2024) doi:10.1002/jmri.29446.
35. Dai, L. *et al.* A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nat. Commun.* **12**, 3242 (2021).
36. Krittanawong, C. Future Physicians in the Era of Precision Cardiovascular Medicine. *Circulation* **136**, 1572–1574 (2017).
37. Lau, E. & Wu, J. C. Omics, Big Data, and Precision Medicine in Cardiovascular Sciences. *Circ. Res.* **122**, 1165–1168 (2018).
38. Dziopa, K., Chaturvedi, N., Asselbergs, F. W. & Schmidt, A. F. Identifying and ranking non-traditional risk factors for cardiovascular disease prediction in people with type 2 diabetes. *Commun. Med.* **5**, 77 (2025).
39. DeGroat, W. *et al.* Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine. *Sci. Rep.* **14**, 1 (2024).
40. Eastwood, S. V. *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLOS ONE* **11**, e0162388 (2016).
41. Burton, M. J. *et al.* The Lancet Global Health Commission on Global Eye Health: vision beyond 2020. *Lancet Glob. Health* **9**, e489–e551 (2021).
42. Flaxman, S. R. *et al.* Global causes of blindness and distance vision impairment 1990-2020: a systematic review and meta-analysis. *Lancet Glob. Health* **5**, e1221–e1234 (2017).
43. Loewenstein, A. *et al.* Save our Sight (SOS): a collective call-to-action for enhanced retinal care across health systems in high income countries. *Eye* **37**, 3351–3359 (2023).
44. Ye, X. *et al.* Osteocalcin and Risks of Incident Diabetes and Diabetic Kidney Disease: A 4.6-Year Prospective Cohort Study. *Diabetes Care* **45**, 830–836 (2022).
45. Tang, Z. *et al.* Relationship of OCT-Based Diabetic Retinal Neurodegeneration to the Development and Progression of Diabetic Retinopathy: A Cohort Study. *Invest. Ophthalmol. Vis. Sci.* **66**, 32 (2025).
46. Majithia, S. *et al.* Cohort Profile: The Singapore Epidemiology of Eye Diseases study (SEED). *Int. J. Epidemiol.* **50**, 41–52 (2021).
47. Thinggaard, B. S. *et al.* The I-OPTA Questionnaire: A National Assessment of Patients with Neovascular Age-Related Macular Degeneration. *Ophthalmol. Ther.* **13**, 3035–3046 (2024).
48. Vitale, S. *et al.* Association of 2-Year Progression Along the AREDS AMD Scale and Development of Late Age-Related Macular Degeneration or Loss of Visual Acuity: AREDS Report 41. *JAMA Ophthalmol.* **138**, 610 (2020).
49. Fu, H. *et al.* Evaluation of Retinal Image Quality Assessment Networks in Different Color-Spaces. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (eds Shen, D. et al.) 48–56 (Springer International Publishing, Cham, 2019).
50. Zhou, J. *et al.* iBOT: Image BERT Pre-Training with Online Tokenizer. Preprint at https://doi.org/10.48550/arXiv.2111.07832 (2022).
51. Grill, J.-B. *et al.* Bootstrap your own latent: A new approach to self-supervised Learning. Preprint at https://doi.org/10.48550/arXiv.2006.07733 (2020).

52. Zhang, H. *et al.* DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. Preprint at https://doi.org/10.48550/arXiv.2203.03605 (2022).
53. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. Preprint at https://doi.org/10.48550/arXiv.1711.05101 (2019).
54. Popescu, D. M. *et al.* Arrhythmic sudden death survival prediction using deep learning analysis of scarring in the heart. *Nat. Cardiovasc. Res.* **1**, 334–343 (2022).

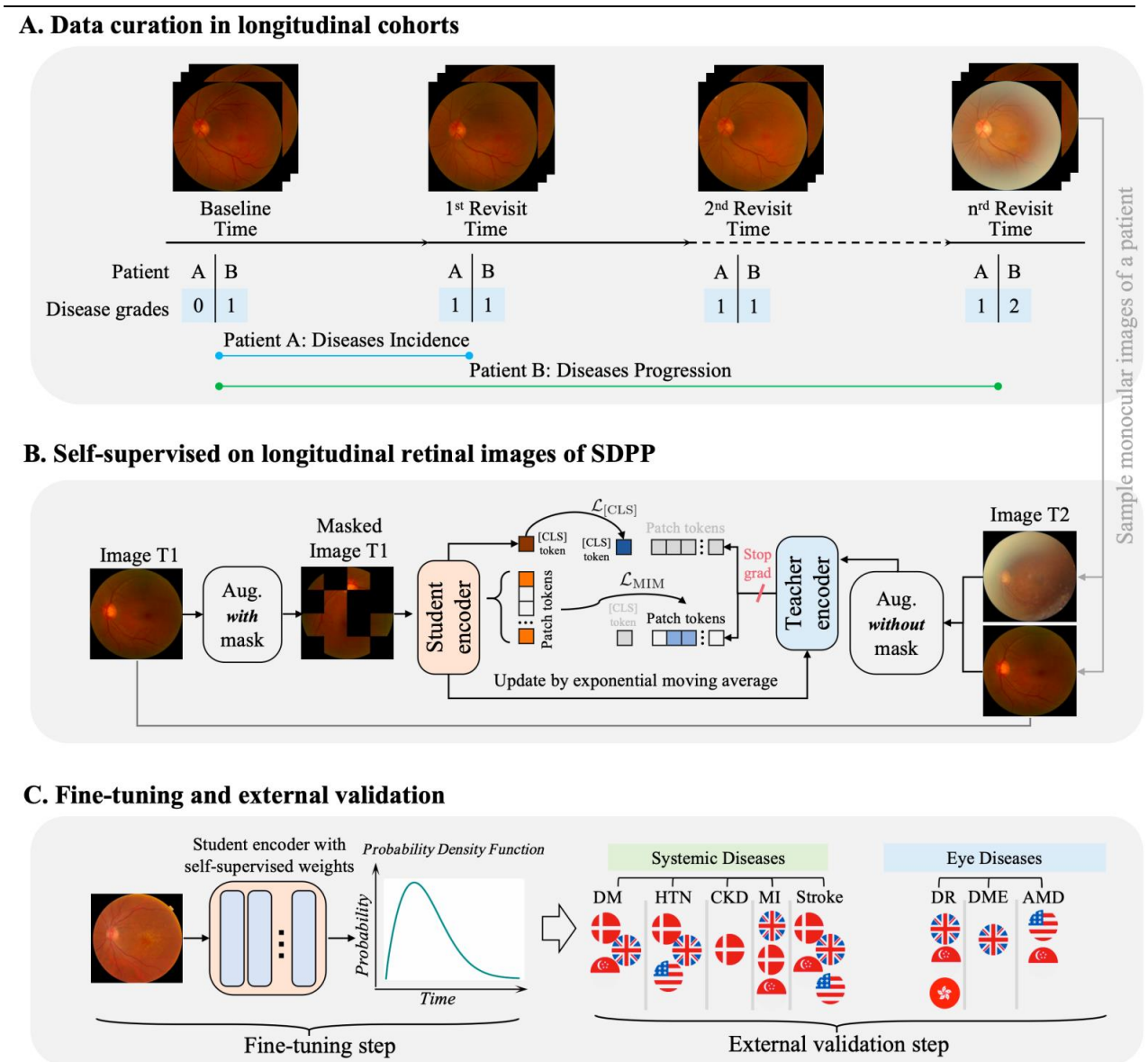
A. Data curation in longitudinal cohorts
Baseline Time
1st Revisit Time
2nd Revisit Time
nrd Revisit Time
Patient
Disease grades
Patient A: Diseases Incidence
Patient B: Diseases Progression
Sample monocular images of a patient
B. Self-supervised on longitudinal retinal images of SDPP
Image T1
Aug. with mask
Masked Image T1
Student encoder
[CLS] token
Patch tokens
Stop grad
Teacher encoder
Aug. without mask
Image T2
Update by exponential moving average
C. Fine-tuning and external validation
Student encoder with self-supervised weights
Probability Density Function
Probability
Time
Fine-tuning step
Systemic Diseases
DM
HTN
CKD
MI
Stroke
Eye Diseases
DR
DME
AMD
External validation step

Figure 1

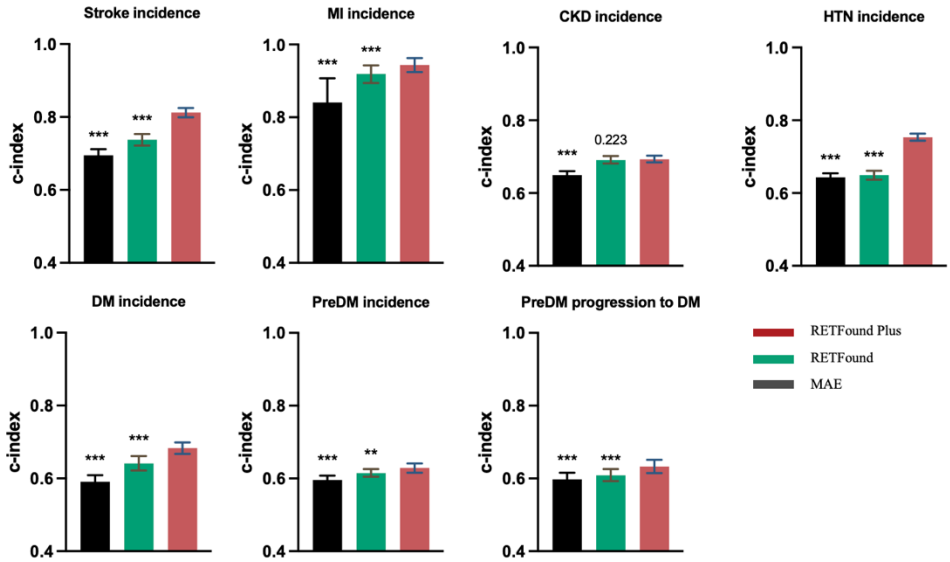

Figure 2

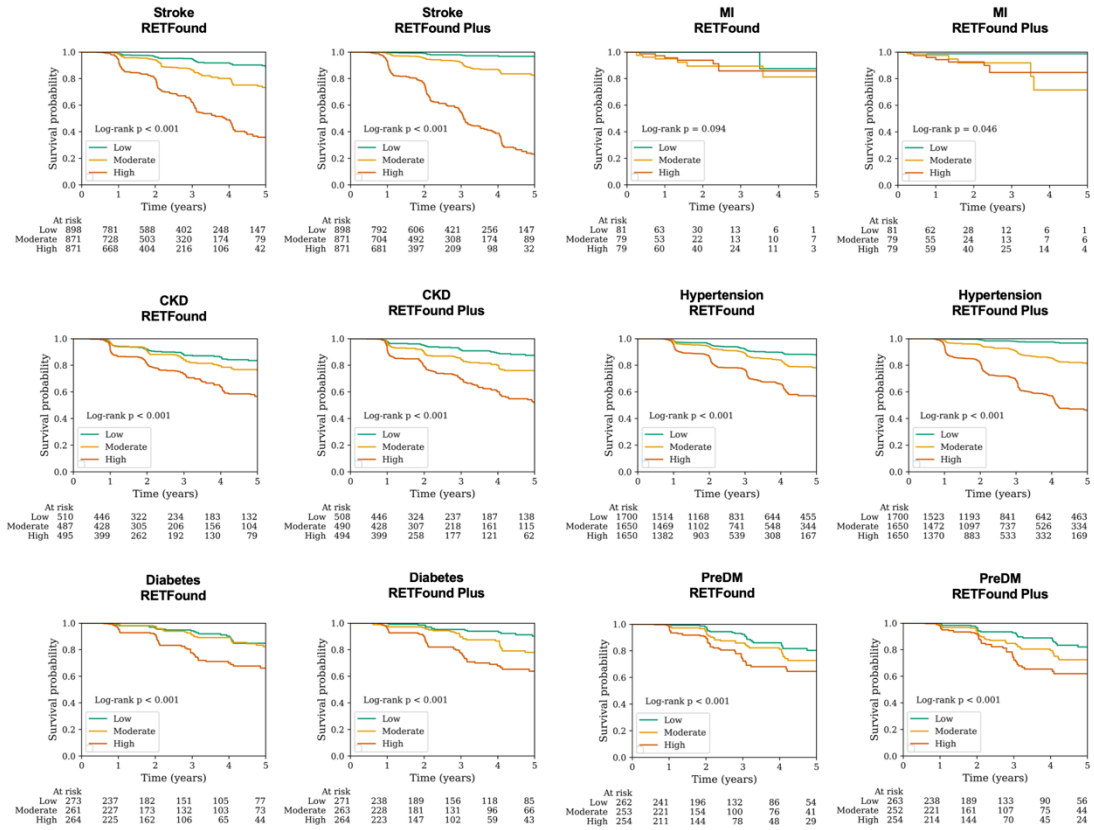

Figure 3

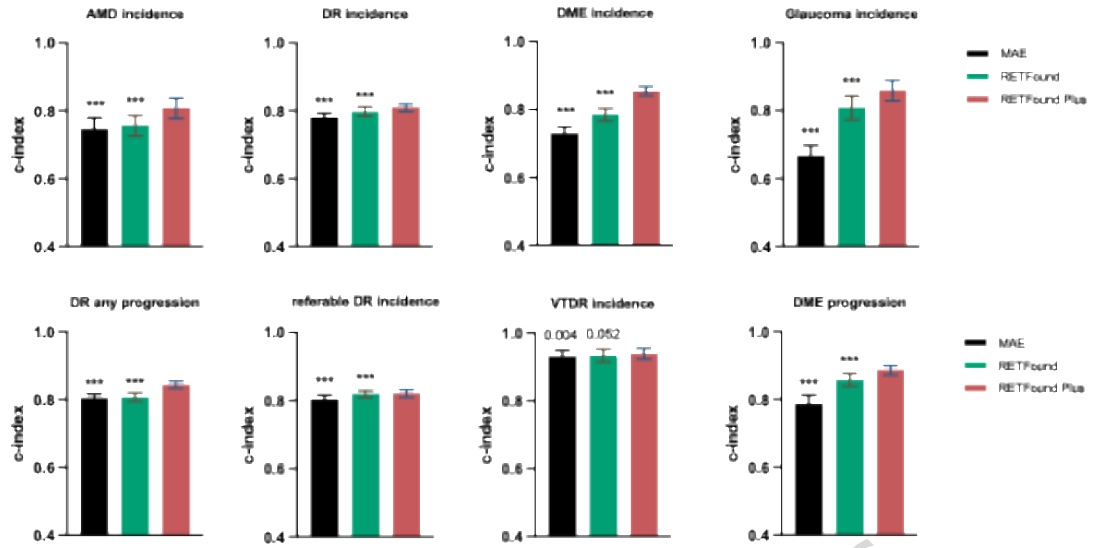

Figure 4

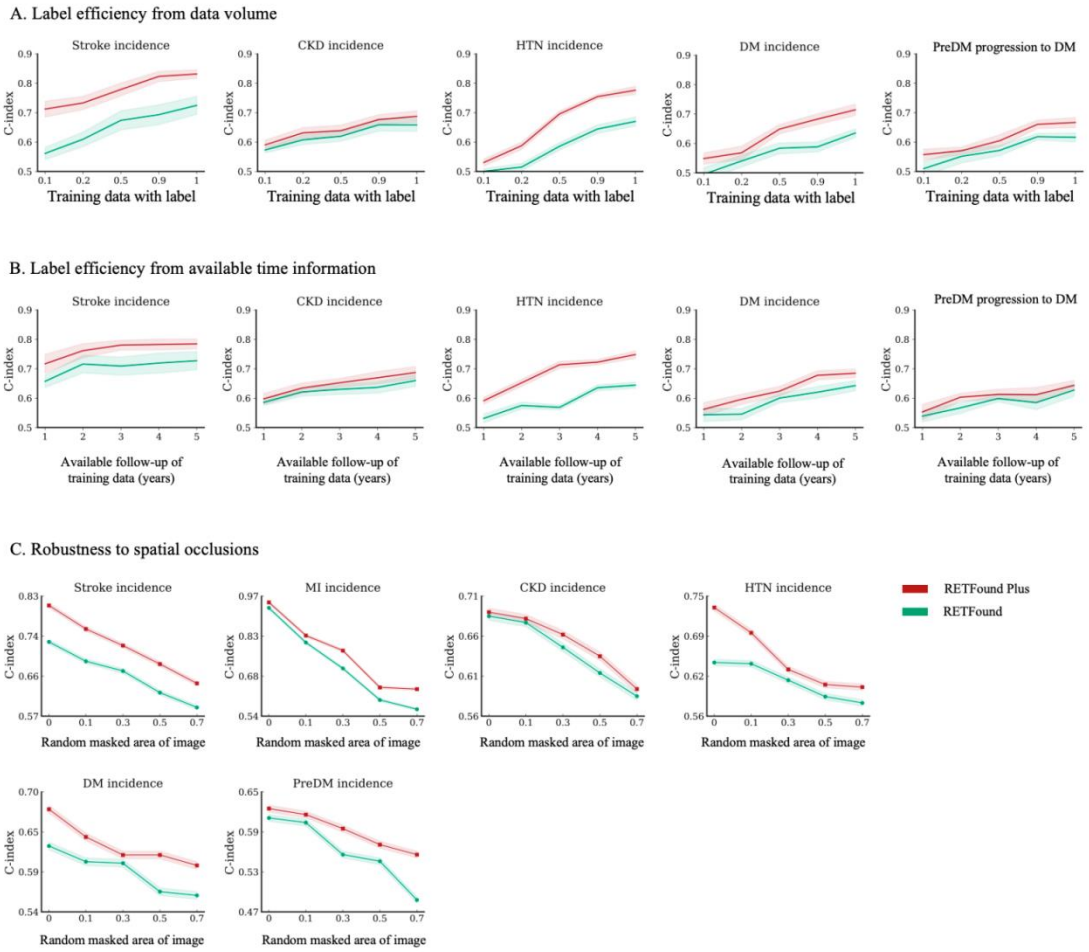

Figure 5