nature medicine

Article

https://doi.org/10.1038/s41591-025-04006-w

A full life cycle biological clock based on routine clinical data and its impact in health and diseases

Received: 6 April 2025

Accepted: 11 September 2025

Published online: 27 October 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Aging research has primarily focused on adult aging clocks, leaving a critical gap in understanding a biological clock across the full life cycle, particularly during infancy and childhood. Here we introduce LifeClock, a biological clock model that predicts biological age across all life stages using routine electronic health records and laboratory test data. To enhance individualized predictions, we integrated virtual patient representations from 24,633,025 heterogeneous longitudinal clinical visits across 9,680,764 individuals and projected them into a latent space. Our approach leverages EHRFormer, a time-series transformer-based model, to analyze developmental and aging dynamics with high precision and develop accurate biological age clocks spanning infancy to old age. Our findings reveal distinct biological clock patterns across different life stages. The pediatric clock is strongly associated with children's development and accurately predicts current and future risks of major pediatric diseases, including malnutrition, growth and developmental abnormalities. The adult clock is strongly associated with aging and accurately predicts current and future risks of major age-related diseases, such as diabetes, renal failure, stroke and cardiovascular diseases. This work therefore distinguishes pediatric development from adult aging, establishing a novel framework to advance precision health by leveraging routine clinical data across the entire lifespan.

Aging is a complex, multifaceted process involving molecular, cellular and organ-level changes that ultimately impact whole-organism health and survival[1]. Understanding how these changes contribute to increased disease susceptibility is essential for developing interventions that extend healthspan[2,3]. Biological age (BA), a measure of accumulated biological damage relative to an average individual of the same chronological age (CA), has emerged as a key metric for assessing age-related disease risk[1]. BA can diverge from CA, providing a valuable indicator of aging trajectories and health outcomes[4].

Initially, BA estimation relied on the measurement of DNA methylation and transcription patterns[2,5], but recent advancements have expanded aging clocks to incorporate imaging and multi-omics data, improving the accuracy and comprehensiveness of BA predictions[6–8]. For example, mass spectrometry and antibody-based proteomics and metabomics have enabled large-scale serum analyses, generating valuable resources for aging research[9]. Furthermore, various medical images and electronic health record (EHR) modalities have provided organ functional aging assessments and linkage to health and diseases[10–14]. These innovations highlight the variability in aging across organs and their differing responses to external factors such as lifestyle or medications, paving the way for personalized anti-aging strategies[15,16].

e-mail: xmchen301@126.com; kang.zhang@gmail.com

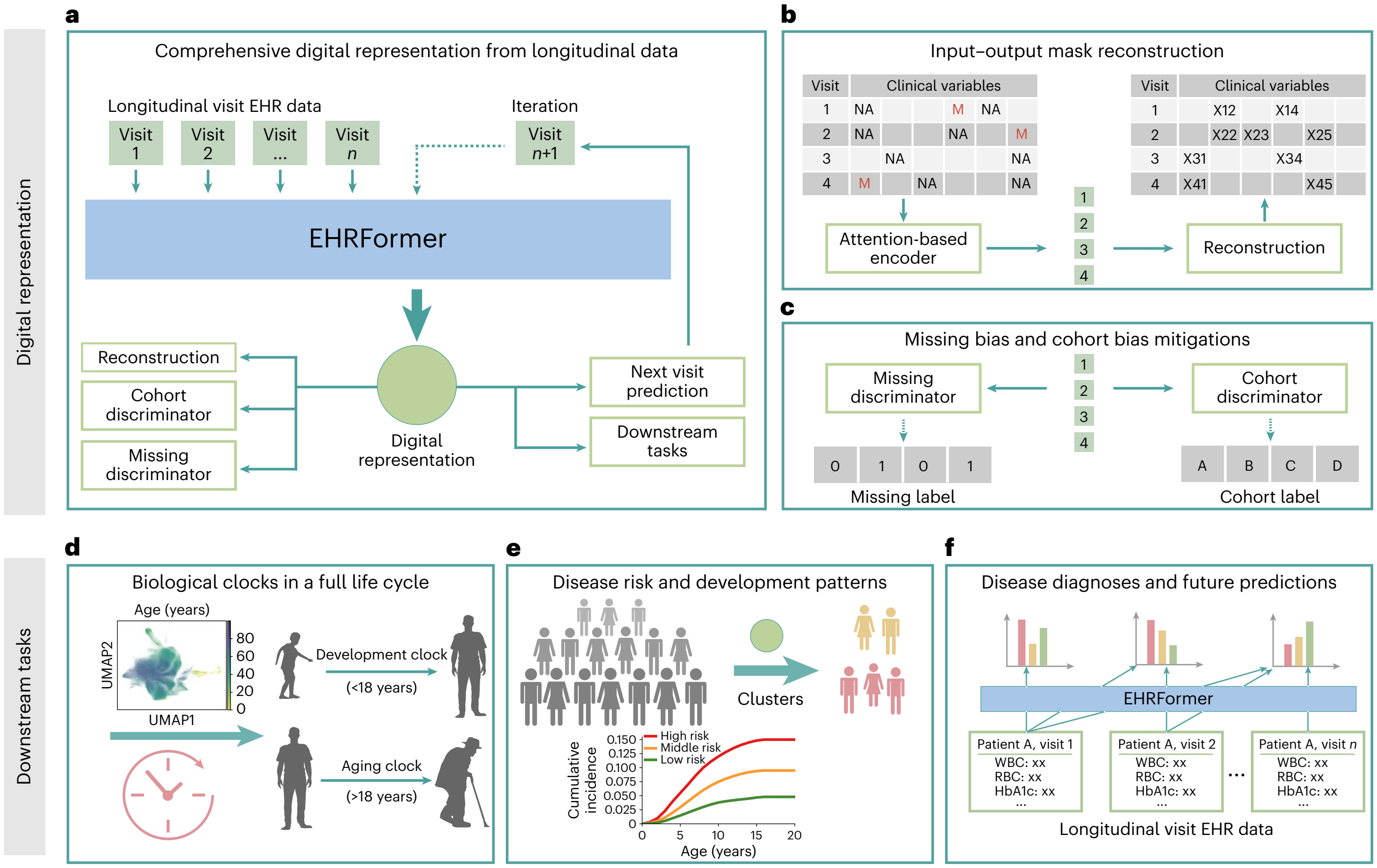


Fig. 1 | EHRFormer architecture and applications for longitudinal EHR data analysis. a, Computational framework for generating comprehensive digital representations from sequential EHR data using EHRFormer. Multiple self-supervised pretraining tasks (reconstruction, cohort discrimination, missing data discrimination and next-visit prediction) optimize the model for EHR feature extraction. **b**, Input–output mask reconstruction mechanism with attention-based encoding for handling masked and missing values. **c**, Bias-mitigation strategies employing discriminators to address missing data patterns and cohort-specific variations. **d**, Biological developmental clock for developmental processes (<18 years) and biological aging clock for aging processes (>18 years) across the human lifespan. **e**, Disease risk stratification and temporal development patterns derived from patient clustering, illustrating differential risk trajectories. **f**, Clinical application diagram depicting first occurrence disease diagnosis and future disease prediction based on longitudinal patient visit data using EHRFormer.

The growing interest in aging mechanisms and interventions has driven interest in aging clocks—molecular markers that predict BA more precisely than CA, which measures the passage of time. Unlike CA, which is static, BA reflects the efficiency of biological functions using genomic, epigenetic, clinical and functional markers[17,18]. Genomic markers are fixed at birth, whereas epigenetic markers, such as DNA methylation and histone modifications, change with age[4,19,20].

In theory, individuals of the same CA should exhibit similar rates of functional decline. However, genetic and environmental factors influence cellular, tissue and organ aging, making some individuals age more quickly or slowly biologically compared to their CA. This discrepancy, quantified as the difference between predicted BA and CA, is known as the age gap[21]. Studies have shown that an increased age gap is associated with accelerated aging and heightened disease risk and mortality[10–14]. For example, individuals with increased brain age gaps often exhibit systemic aging features, such as sensory-motor decline and older appearance[22]. Accelerated aging is particularly evident in individuals with chronic diseases, suggesting that disease burden further drives biological aging[23]. Conversely, accelerated biological aging may also serve as a strong determining factor in shaping disease risk before onset, as demonstrated through incident disease and multimorbidity analyses[24]. By developing reliable BA measures, aging clocks hold promise for extending healthspan and improving quality of life, a crucial goal as human life expectancy continues to rise[25].

Despite substantial progress in adult aging clocks, our understanding of a full life cycle clock, particularly during infancy and childhood, and its impact on health and disease remains limited[15,19,20,26]. In pediatrics, rapid child physiological changes represent a scripted development progression rather than accumulated biological aging damage, making the current BA definition ill-posed in the pediatric context, in which the term 'physiological maturity clock' may be more appropriate. Furthermore, a study on the link between (and relevance of) clock deviations and clinical implications would have great potential in pediatric care. For example, a desirable goal is to calculate 'physiological maturity' deviations either in maturation precocity or puberty precocity versus growth/developmental delays/malnutrition relative to peers. This information may provide a useful clinical interpretation and facilitate pediatric care and interventions in the setting of pediatric growth charts or preventive screening programs.

This study introduces LifeClock, a full life cycle biological clock leveraging 24.6 million EHRs, including laboratory test data, to predict BA across all life stages and assess its association with disease risk and survival outcomes. Physicians traditionally focus on EHR indicators/laboratory values that exceed reference ranges, yet normal values also contain valuable insights. Integrating longitudinal data—regardless of whether values are normal or abnormal—can help identify individual-specific setpoints and fluctuations, improving disease risk assessment and the detection of critical aging transitions[25]. Although

Table 1 | Demographic characteristics of the study cohorts

	Cohort #1	Cohort #2	Cohort #3	Cohort #4	Cohort #5 (External)	UK Biobank
≤18 years	732,920 (4.87%)	2,938,942 (47.06%)	74,808 (8.31%)	960,015 (39.22%)	41,937 (19.11%)	–
18–30 years	1,960,491 (13.03%)	703,787 (11.27%)	57,640 (6.41%)	641,319 (26.21%)	29,967 (13.65%)	–
30–50 years	5,508,508 (36.62%)	1,275,869 (20.43%)	178,843 (19.87%)	830,256 (33.93%)	69,443 (31.63%)	35,916 (21.60%)
50–70 years	5,180,385 (34.44%)	940,200 (15.06%)	347,075 (38.56%)	14,767 (0.6%)	57,759 (26.32%)	126,808 (76.24%)
≥70 years	1,658,228 (11.03%)	386,224 (6.18%)	241,497 (26.84%)	1,251 (0.05%)	20,379 (9.28%)	2,721 (1.64%)
No. of records	15,040,532	6,245,022	899,863	2,447,608	219,485	166,317
Male	6,823,323 (45.37%)	3,158,310 (50.57%)	497,942 (55.33%)	575,429 (23.51%)	107,412 (48.94%)	75,939 (45.67%)
Female	8,217,209 (54.63%)	3,086,712 (49.43%)	401,921 (44.66%)	1,872,179 (76.49%)	112,073 (51.06%)	90,378 (54.33%)
No. of participants	4,623,404	4,022,651	437,396	597,313	86,257	146,087
Male	2,175,316 (47.05%)	2,048,310 (50.92%)	241,573 (55.23%)	208,539 (34.91%)	41,644 (48.28%)	66,032 (45.21%)
Female	2,448,088 (52.95%)	1,974,341 (49.08%)	195,823 (44.77%)	388,774 (65.09%)	44,613 (51.72%)	80,055 (54.79%)

This table displays the distribution of participant records by age group and sex across ten different hospital cohorts used for model development and validation. Cohorts #1, #2, #3 and #4 were utilized for model development and internal validation. Cohorts #5 and UK Biobank were utilized for external validation, primarily comprising normal individuals (healthy controls). For model development and internal validation, data were aggregated from four hospital cohorts (cohorts #1, #2, #3 and #4). This combined training and internal validation dataset comprised a total of 9,680,764 unique participants, accounting for 24,633,025 records (visits). The external validation set consisted of data from an additional cohort #5. Based on the available summary data presented in the table, this external validation set included 86,257 unique participants and 219,485 visits. The UK Biobank external validation set included 146,087 unique participants and 166,317 visits.

deep learning models have the capacity to extract such information, previous studies have largely focused on specific diseases within narrow age ranges[27–29]. Previous studies on BA models derived from routine clinical labs and vitals (for example, PhenoAge, Klemera-Doubal and DOSI) and EHR were largely focused on single-visit or cross-sectional studies, making it difficult to capture longitudinal trajectories to achieve good predictive performance or clinical interpretability.

To further advance precision aging health research and clinical applications, virtual representations of individual patients[30] were generated using massive (24,633,025) longitudinal EHR data through EHRFormer, a transformer-based model. This approach enabled high-granularity modeling of aging processes, enhancing our understanding of the interplay between biological aging and disease risk, allowing us to identify distinct clinical patterns correlated with age, stratifying individuals into unique clusters with varying disease trajectories.

Using unsupervised learning, we trained EHRFormer to extract features from vast patient data spanning birth and childhood to adulthood and the geriatric phase, to provide more accurate BA estimates than CA. The model integrates EHR data that reflect the functioning of multiple organ systems—including blood, immune, liver and kidney—while also accounting for sex differences in aging patterns. Model performance was evaluated by comparing predicted BA to CA using R^2, the Pearson correlation coefficient (PCC) and mean absolute error (MAE), demonstrating high accuracy, particularly in younger individuals with more uniform developmental trajectories. By focusing on biological aging, our study also identifies age gaps—notable divergences between BA and CA—as critical biomarkers for disease risk prediction and patient stratification. This is especially relevant in cases of accelerated aging, which correlate with increased disease risk across both younger and older populations. Our study presents a novel framework for studying aging and age-related diseases across the full life cycle, leveraging widely available and cost-effective EHR data to advance precision medicine in aging research.

Results

A blood test-based biological clock in a full life cycle using longitudinal EHRs

To construct a virtual representation of human health from rich, longitudinal EHRs, we first had to overcome inherent data challenges such as heterogeneity, missing values and cohort-specific batch effects. To address this, we developed a foundation model, EHRFormer (Fig. 1a), using data from multiple cohorts (Table 1 and Extended Data Fig. 10), starting with 184 carefully selected clinical indicators (Supplementary Table 1) from the China Healthy Aging Investigation (CHAI). The model's architecture incorporates several key strategies: an input–output dual stochastic masking strategy to capture complex feature interactions while imputing missing data (Fig. 1b), and a cohort-agnostic adversarial training model to eliminate batch effects, ensuring the representations are robust and generalizable (Fig. 1c). Furthermore, an autoregressive training approach was used to ensure each visit's representation captures an individual's evolving health trajectory by learning from past and present records to predict the future (Fig. 1a,e,f).

Using EHRFormer, we generated digital representations from each visit of healthy individuals and developed a task-specific regression model to predict CA, with the predicted CA values serving as BA estimates (Fig. 1d). This BA clock demonstrated strong overall performance in the internal validation cohort, achieving a low MAE, high R^2 and high PCC when compared with CA, indicating that laboratory tests alone can reliably estimate CA (Fig. 2a). Our analysis revealed two distinct aging patterns: a pediatric phase (birth to 18 years) and an adult phase (18 years onward), which were characterized by markedly different profiles of laboratory markers (Fig. 2a,b). Consequently, we trained separate, specialized models for each phase, which substantially improved prediction accuracy (Fig. 2c,e).

Visual explanations using 'Shapley additive explanations' (SHAP) identified the key contributors for each clock. The pediatric clock was primarily driven by low aspartate aminotransferase (AST), high creatinine (crea) and high total protein (TP) levels (Fig. 2d). In contrast, the adult clock's most influential features were high urea, low albumin (ALB) and high red cell distribution width (RDW) (Fig. 2f), with the top 20 markers being almost entirely different between the two clocks. The model's performance was consistent across sexes (Extended Data Fig. 1a,c,e,g), though feature contributions varied slightly (Extended Data Fig. 1b,d,f,h). Importantly, EHRFormer's predictive power was validated in the external UK Biobank cohort, where it achieved an MAE of 4.14 (Extended Data Fig. 7a). Key aging biomarkers such as urea, ALB and RDW were identified as top contributors in both the CHAI and UK Biobank cohorts, demonstrating their cross-cohort stability (Extended Data Fig. 7b).

LifeClock predicts current and future disease risks in both children and adults

We applied our EHRFormer-derived representations for dimensionality reduction using principal component analysis (PCA) and uniform

https://doi.org/10.1038/s41591-025-04006-w

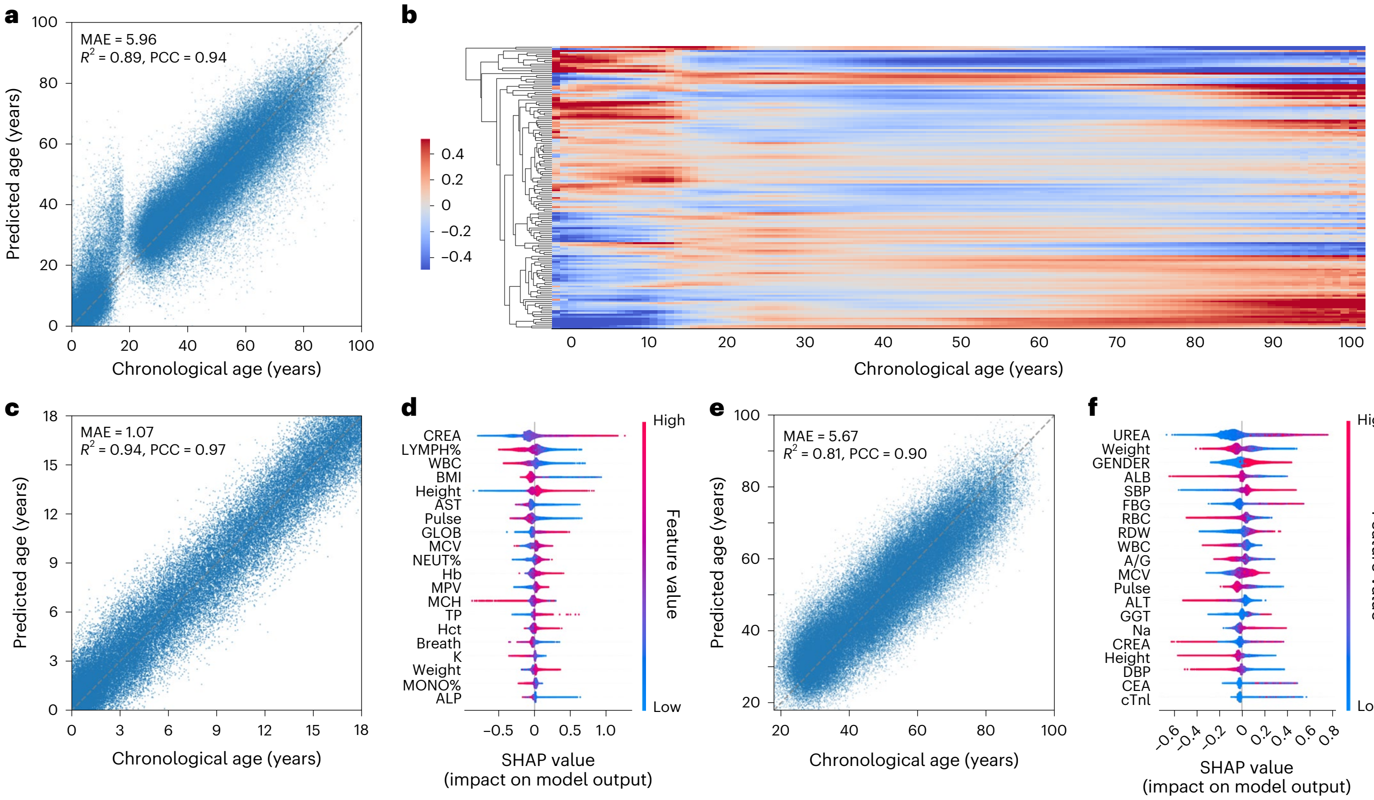


Fig. 2 | Overall BA prediction model, specific models and associated features for predictions on the pediatric development clock and adult aging clock. **a**, The strong correlation between CA and BA predicted by the EHRFormer-based age model. Each dot represents one EHR data point. **b**, Heatmap showing the average values of multiple clinical parameters (*z*-score normalized across the population) clustered by their age-related trajectories, with data aggregated at one-year intervals across the entire study population. **c**, Correlation between CA and BA in the pediatric clock predicted by the EHRFormer-based age model. Each dot represents one EHR data point under 18 years of age. **d**, SHAP values of the top 20 contributors in the BA prediction for EHR data under 18 years of age. **e**, Correlation between CA and BA in the adult clock predicted by the EHRFormer-based age model. Each dot represents one EHR data point over 18 years of age. **f**, SHAP values of the top 20 contributors in the BA prediction for EHR data over 18 years of age.

manifold approximation and projection (UMAP), followed by Leiden clustering analysis[29]. Our results revealed that, among healthy individuals, different CA groups could be clearly clustered, indicating that EHR data contain age-related information (Extended Data Fig. 2j). Furthermore, data from different hospitals or cohorts were evenly distributed across clusters, particularly when separating those under 18 and over 18 years of age, indicating the successful elimination of batch effects (Extended Data Fig. 2).

Given the well-established link between BA and disease risks[31], we performed dimensionality reduction followed by a Leiden clustering analysis on the entire CHAI dataset and examined whether individuals with higher age differences were more likely to develop diseases. We also explored potential associations between the clusters and diseases (Fig. 3 and Supplementary Tables 2 and 3). Our aging model, built on the EHRFormer framework and trained on healthy individuals, computed an age difference for each individual in the CHAI dataset by quantifying deviations from the individual's BA relative to same-CA peers through analysis of EHR profiles (Methods). A total of 64 Leiden clusters were obtained from all EHR representations (Fig. 3b). We classified adult EHRs into two categories, average-aged (age difference within ±1 s.d.) and over-aged (age difference > 3 s.d.), and then calculated the prevalence and incidence proportions of different diseases within each cluster. Our results showed that, for most diseases, a markedly higher disease prevalence proportion was present in over-aged individuals when compared to average-aged individuals within the same cluster, which was further increased in the future (Fig. 3c,d). In addition, some diseases within certain clusters, such as hypoglycemia, may exhibit a higher incidence proportion in the future in the over-aged individuals, even though these over-aged individuals may not have demonstrated a higher prevalence proportion (Fig. 3e). In summary, these results suggest that the EHRFormer-based aging model may not only demonstrate the present health status but also indicate future disease risk based on current EHR profiles.

Because clusters can serve as indicators of future disease risks, and the EHR representations for children (<18 years old, clusters 1–14) were well separated from those for adults (>18 years old, clusters 15–64), we analyzed the disease risks separately within the children (0–20 years old) and adult (>20 years old) clusters, respectively. For each identified cluster, we used Cox proportional hazards models for incidence calculations using the cluster assigned at each patient's first clinical visit as a baseline predictor for the remainder of the study population, applying multivariate adjustment for age, sex, hospital, smoking and alcohol history to minimize potential confounding from demographic factors and institutional variations. Sex was included as a covariate rather than used for stratification to maintain statistical power across all clusters. As a result, in clusters 1–14, by calculating adjusted $\log_2$ hazard ratios (HRs) for incidence (between ages 12 and 20 years, which represent the children maturation period) using EHR data from individuals <12 years of age (before the children maturation period), we observed that individuals within different clusters exhibited distinct tendencies to develop specific pediatric disease conditions. For instance, we found that individuals in cluster 14 had 15.36 times and 11.07 times higher risk of developing pituitary hyperfunction and obesity, respectively; individuals in cluster 12 had a 10.13 times higher risk of developing hernia; individuals in cluster 3 had 4.71 times higher risk of developing viral meningitis; and individuals in cluster 8 had 4.95 times higher risk

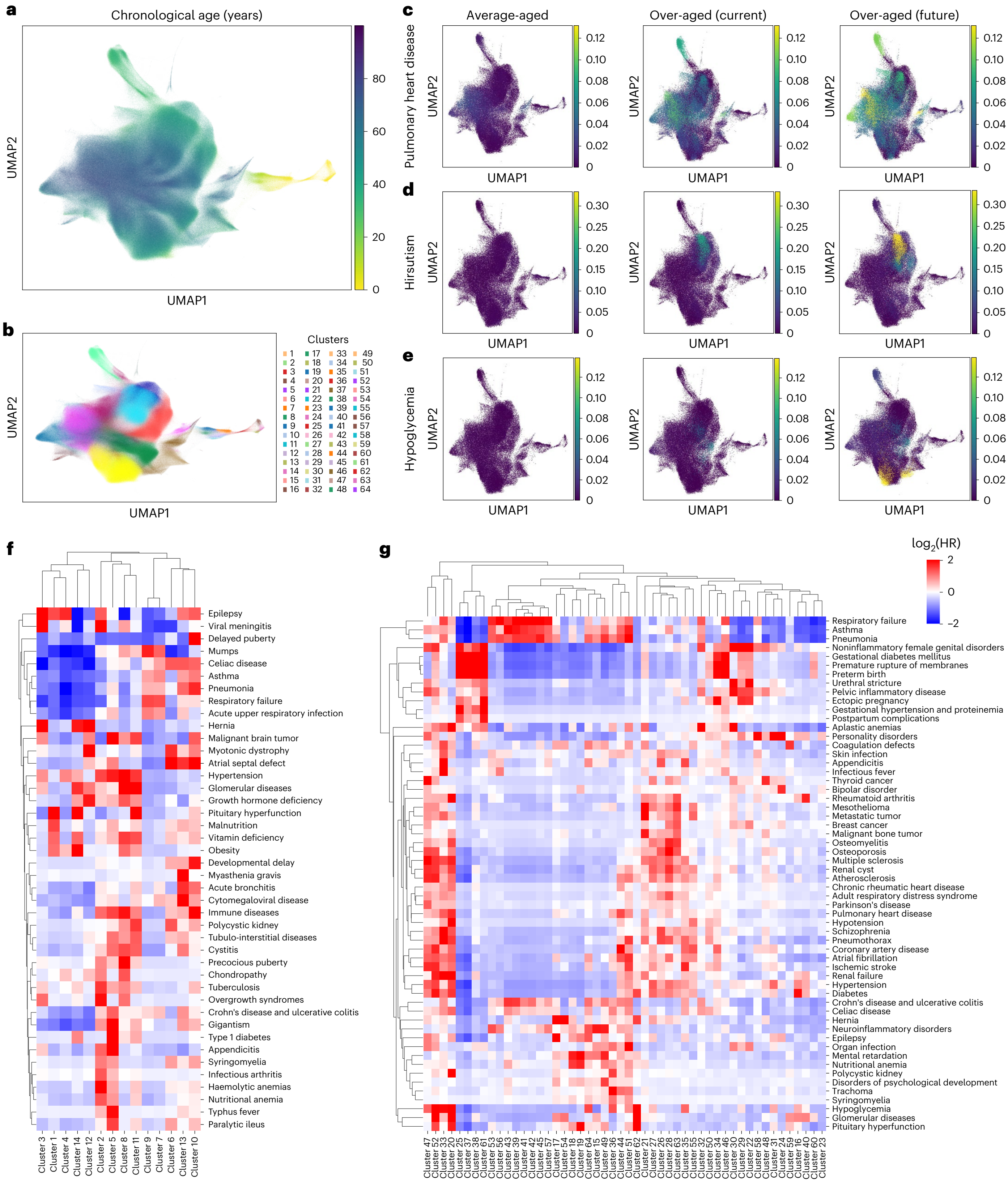


Fig. 3 | Clusters generated based on EHRFormer-representations are informative of current and future health status. a, UMAP projection of patient visits colored by age (yellow to purple: young to old). Each point represents one EHR data point (one patient visit). **b**, Leiden cluster distribution of the UMAP in **a**. **c**–**e**, The prevalence proportion of pulmonary heart disease (**c**), hirsutism (**d**) and hypoglycemia (**e**) in the average-aged group (left), the current over-aged group (middle) and the future over-aged group (right). **f**, Heatmap showing the adjusted $\log_2$HRs of 42 diseases in each of the 14 clusters of <12-year-old pediatric EHR data. **g**, Heatmap showing the $\log_2$(HR) of 58 common diseases in each of the 50 clusters using >18-year-old adult EHR data. Col-or from blue to red indicates the $\log_2$(HR) in a specific cluster truncated at a maximum absolute value of 2.

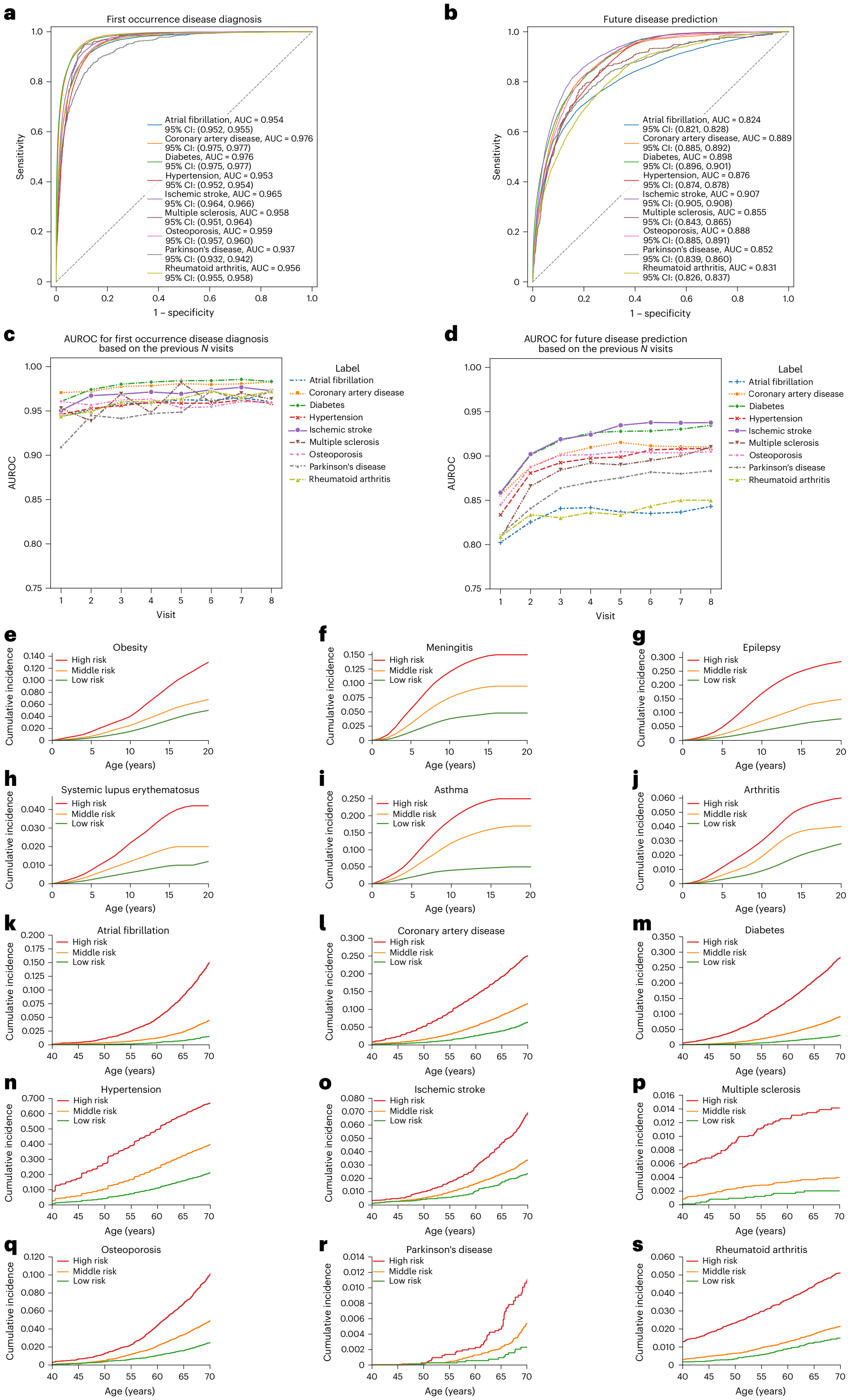


Fig. 4 | Performance of the EHRFormer-based disease predicting model and accumulated risk analysis in the CHAI-Internal cohort. a, ROC curves of the EHRFormer-based disease predicting model in diagnosing different diseases. **b**, ROC curves of the EHRFormer-based disease prediction model in predicting future risk of different diseases. **c**, AUC changes along with visit number in diagnosing different diseases. **d**, AUC changes along with visit number in predicting different diseases in the future. **e**–**j**, Accumulated risk for various diseases based on a predictive model that categorized individuals into high-, middle- and low-risk groups for each given age under 20 years. As development progresses, the gap in accumulated risk between the high-, middle- and low-risk groups becomes more pronounced. **k**–**s**, Accumulated risk for various diseases based on a predictive model that categorized individuals into high-, middle- and low-risk groups for each given age over 40 years. As aging progresses, the gap in accumulated risk between the high-, middle- and low-risk groups becomes more pronounced.

of developing precocity puberty. In contrast, individuals in cluster 10 had 3.57 times higher risk of developing developmental growth delay (Fig. 3f). Furthermore, analysis of developmental clock-derived age differences in children <18 years showed significant BA deceleration in growth-inhibiting conditions (delayed puberty, growth hormone deficiency and developmental delay) compared to healthy controls. Conversely, growth-promoting disorders (precocious puberty, gigantism and overgrowth syndromes) exhibited marked BA acceleration, demonstrating that our developmental clock captures physiologically meaningful growth variations (Extended Data Fig. 4).

In parallel, within clusters 15–64, individuals in cluster 20 had a more than 30 times increased risk of vascular-related disorders, including hypotension (9.03 times) and renal failure (37.70 times) (Fig. 3e). Similarly, diabetes showed increased HRs for incidence by 3.75, 3.59 and 3.00 times in clusters 16, 52 and 20, respectively (Fig. 3e). These findings demonstrate that our model can effectively identify individuals at high risk of developing diseases based on their longitudinal EHR data. To further interpret these high-risk clusters, we examined their underlying clinical profiles. For instance, in the pediatric cohort, cluster 5, which showed a higher incidence of appendicitis, ulcerative colitis and other immune-related diseases (Fig. 3f), was correspondingly characterized by elevated immune and inflammatory markers, including interleukin-6 (IL-6), IL-8, IL-10, white blood cell count (WBC) and C-reactive protein (CRP) (Extended Data Fig. 3a). Similarly, in the adult cohort, cluster 44 was associated with a substantially higher incidence of cardiopulmonary diseases (Fig. 3g) and was defined by a corresponding clinical signature of elevated cardiac troponin T (cTnT) and serum potassium, alongside lower oxygen saturation (saO_2) (Extended Data Fig. 3b).

Fine-tuning EHRFormer for individual disease risk predictions

Because the success of the EHRFormer-based BA prediction model in indicating current disease diagnosis and future disease predictions suggests that EHR data may contain information beyond aging, such as overall health status and disease progression, we speculated that our EHRFormer model could be fine-tuned with the introduction of disease labels for disease risk predictions. This approach would enhance the model's ability to diagnose first occurrence disease status and predict future disease, enabling a quantitative assessment of its predictive capabilities (Fig. 1f). For each predicted disease, we stratified the population into high-, middle- and low-risk cohorts based on model-generated probability scores, then quantified cumulative risk profiles across age groups. This stratification approach enables age-specific risk assessment and potentially facilitates the identification of critical intervention windows within the disease trajectory (Fig. 1e).

We found that our EHRFormer-based disease prediction model demonstrated strong current diagnostic performance across multiple diseases. Specifically, it achieved a high prediction accuracy in cardiovascular diseases (atrial fibrillation area under the curve (AUC) = 0.95, coronary artery disease (CAD) AUC = 0.98, hypertension AUC = 0.95, ischemic stroke AUC = 0.97), neurological disorders (multiple sclerosis AUC = 0.96, Parkinson's AUC = 0.94) and systemic conditions (osteoporosis AUC = 0.96, rheumatoid arthritis AUC = 0.96, diabetes AUC = 0.98) (Fig. 4a and Supplementary Table 4). Additionally, the model effectively predicted future risks of these diseases (AUC ≥ 0.8) (Fig. 4b and Supplementary Table 4). To further assess its capability for long-term risk stratification, we specifically evaluated its performance on five-year and ten-year incidence-prediction tasks. The model maintained strong predictive power, achieving AUCs ranging from 0.80 to 0.90 for five-year incidence (Extended Data Fig. 5a) and 0.81 to 0.91 for ten-year incidence across various diseases (Extended Data Fig. 5b). For comparison, we also evaluated EHRFormer against other models such as Recurrent Neural Network (RNN) and XGBoost. RNN, similar to EHRFormer, accepts sequential data and follows the autoregressive paradigm, but lacks an attention mechanism. XGBoost can also handle sequential data, yet it does not operate under the autoregressive framework. EHRFormer also demonstrated a superior performance compared to XGBoost and RNN in current disease diagnosis tasks across nine diseases. For example, in atrial fibrillation diagnosis, EHRFormer achieved an area under the receiver operating characteristic curve (AUROC) of 0.962, while XGBoost achieved 0.899 and RNN 0.907. In diabetes future prediction, EHRFormer's AUROC was 0.911, versus 0.837 for XGBoost and 0.876 for RNN (Supplementary Table 5). To further validate its predictive abilities, we tested the model on an external validation cohort consisting of 219,485 longitudinal clinical visits from 86,257 individuals collected in an independent cohort (Table 1). We fine-tuned the model using each individual hospital's EHR data in CHAI-Training and observed consistently good predictive performance in for CHAI-External cohort #5 (Extended Data Fig. 6). Our model demonstrates robust performance when evaluated on UK Biobank EHRs, comparable to results observed in the CHAI-External cohort, highlighting its high generalizability and consistent effectiveness across diverse populations and healthcare institutions (Extended Data Figs. 7c,d and 8 and Supplementary Table 6). Therefore, by benchmarking against baseline models and analyzing the correlation between the number of visits and predicting accuracy, EHRFormer markedly improved current and future disease prediction performance by integrating the whole life cycle of aging and disease information (Fig. 4c,d).

We also applied the model for future disease risk predictions in both pediatric populations (using EHR data before <12 years of age) and adult populations (using EHR data >18 years of age). Using EHR data collected before the age of 12 years, we predicted future common pediatric disease risks, achieving AUCs ranging from 0.70 to 0.96. Similarly, using EHR data from individuals over 18 years, we predicted future adult diseases with comparable accuracy (Extended Data Fig. 9). Furthermore, we stratified individuals under 10 years old into three risk-level groups based on their predicted probabilities: the highest one-third as the high-risk group, the middle one-third as the medium-risk group and the bottom one-third as the low-risk group. Cumulative incidence plots provide a useful visual tool for comparing disease incidence over time among these groups, revealing large differences in future disease risk for various conditions, including obesity, meningitis, epilepsy, systemic lupus erythematosus (SLE), asthma and juvenile arthritis (Fig. 4e–j). Similarly, we applied the same stratification approach to individuals over 40 years of age, dividing them into three risk-level groups based on predicted probabilities. The cumulative incidence curves for these groups demonstrated substantial differences in future disease risk for atrial fibrillation, coronary artery disease, diabetes, hypertension, ischemic stroke, multiple sclerosis, osteoporosis, Parkinson's disease and rheumatoid arthritis after age 40 (Fig. 4k–s). These results suggest that risk stratification based on early-life pediatric EHR data and early-adulthood EHR data can effectively reveal differential long-term disease risks.

Discussion

This study highlights the potential of EHRFormer as a powerful tool for predicting BA across the full life cycle, providing novel insights into aging processes and their association with disease risks[30,32,33]. By leveraging a large longitudinal cohort of EHR data, our results reveal distinct biological aging clocks in the pediatric and adult phases and demonstrate how deviations from CA—captured as differences between it and predicted BA—are linked to disease susceptibility. These insights offer a unique opportunity to enhance our understanding of aging across the lifespan[29,34].

Building on this foundation, our initial finding of a strong correlation between BA and CA using EHR data (Fig. 2a,c,e) led us to discover age-correlated changes in 184 clinical laboratory test results, vital sign indicators and basic metadata (Fig. 3b,d,e). These features were subsequently subject to a clustering analysis similar to that in single-cell analysis methods (Fig. 3a). The resultant 64 clusters displayed distinct

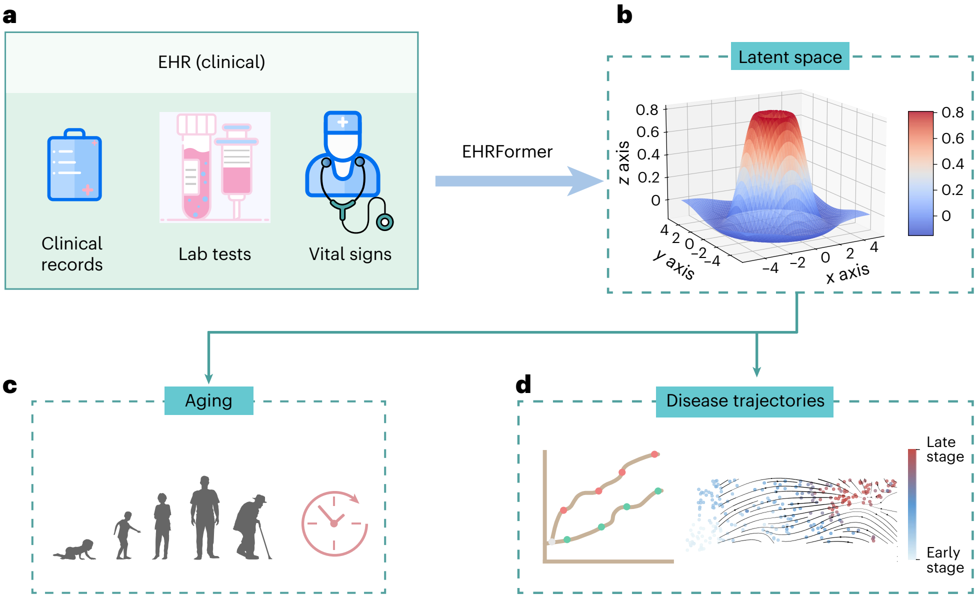


Fig. 5 | A latent space approach for modeling a full life cycle biological clock using the EHRFormer architecture. a, The model takes high-dimensional, longitudinal EHR data as input, including structured clinical records, laboratory test results and vital signs. **b**, EHRFormer, a transformer-based foundation model, processes these data and projects them onto a low-dimensional latent space. This creates a comprehensive and robust 'digital representation' of an individual's health status over time. This learned representation is then leveraged for two key downstream applications. **c**, Modeling biological aging across the full human life cycle, from infancy and childhood development to adult and geriatric aging. **d**, Predicting individual disease trajectories, enabling risk stratification, and identifying patient subgroups with distinct clinical paths.

age characteristics and disease features, which were then subjected to aging assessment and disease risk predictions (Figs. 3 and 4).

The ability to deconstruct a heterogeneous patient population into these clinically meaningful subgroups via unsupervised clustering is a key finding of our study, moving beyond simple disease labels. For example, our analysis revealed that cluster 5 was associated with a high risk for immune-related diseases such as appendicitis in the pediatric cohort and was characterized by elevated inflammatory markers such as IL-6 and CRP (Extended Data Fig. 3a). Given that IL-6 and CRP are canonical biomarkers of systemic inflammation cited in countless studies on pediatric inflammatory conditions[35], our interpretation that cluster 5 represents a state of 'heightened pediatric immune activity or dysregulation' is strongly supported. Similarly, in the adult cohort, cluster 44, which predicted a high incidence of cardiopulmonary diseases, was defined by elevated cardiac troponin T and lower oxygen saturation (Extended Data Fig. 3b), identifying a subpopulation with subclinical or overt cardiorespiratory stress. Furthermore, clusters like cluster 20, with its strong association with renal failure and diabetes, likely represent a well-described metabolic syndrome or vasculopathy phenotype[36]. In the pediatric population, clusters successfully stratified individuals along a spectrum of endocrine and developmental trajectories, capturing conditions from precocious puberty (cluster 8) to developmental delay (cluster 10), reflecting known endocrine feedback loops that govern growth[37]. This mechanistic interpretation of clusters transforms them from abstract groupings into actionable clinical phenotypes that reflect underlying biological states. Notably, although core metabolic markers such as glucose and HbA1c were not top global predictors in our SHAP analysis, their predictive importance was still considerable, likely because our full life cycle model captures metabolic health through a complex interplay of correlated longitudinal markers rather than single-point indicators.

Our work also fits into a broader landscape of foundation models developed for healthcare, unlike models such as OMICmAge[38], which rely on specialized and costly multi-omics data for an aging clock construction. Other models like COMET[29] leverage EHR data through supervised pretraining to enhance the analysis of separate omics datasets. In contrast, EHRFormer demonstrates strong predictive performance using only widely available, low-cost routine laboratory tests and EHR data, enhancing its potential for broad clinical translations, and employs large-scale self-supervised pretraining directly on longitudinal EHRs to learn deep, clinically relevant patient representations without the need for labeled data. Although models such as MILTON[39] excel at integrating unstructured clinical text with structured EHR data, the unique strength of EHRFormer lies in its autoregressive architecture, specifically designed to capture the temporal dynamics and long-range dependencies within an individual's full life cycle and project the information onto a latent space, facilitating aging and age-related disease trajectories (Fig. 5). Therefore, EHRFormer carves a unique niche by focusing on deriving actionable, longitudinal health insights directly from routine clinical data.

Despite its strengths, our model has limitations, including the observational nature of our datasets and potential biases inherent in longitudinal cohorts. Nevertheless, our study underscores the effectiveness of EHRFormer as a virtual representation technology capable of capturing critical health information and providing a novel framework for leveraging widely available EHR data[28]. The strong predictive performance of our representation-based aging clock highlights its potential for future applications in aging research[40–44]. Although traditional aging clocks estimate BA based on specific biomarkers, EHRFormer extends this capability by integrating diverse data sources, offering a dynamic and holistic approach to aging analysis[10,45,46]. By continuously updating with new information, EHRFormer transforms aging clocks from static estimators into adaptive, real-time systems[47–51]. Looking ahead, incorporating wearable devices, cloud medical records and environmental sensors can enable aging clocks to use the most current data, improving their adaptability and accuracy[32,52–54]. The EHRFormer-based clock establishes a robust framework for advancing personalized healthcare strategies, promoting healthy aging, facilitating timely interventions and mitigating aging-related decline.

The findings from this study suggest that our full lifespan aging clock, EHRFormer, offers greater accuracy in predicting disease risk compared to CA alone. The integration of longitudinal EHR data into biological aging models holds the potential to revolutionize our understanding of aging and its relationship with disease. These insights can drive the development of more precise aging biomarkers, enable prompt disease detection, and guide personalized treatments tailored to unique aging trajectories in diverse populations.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-025-04006-w.

References

1. Argentieri, M. A. et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nat. Med.* **30**, 2450–2460 (2024).
2. Bell, C. G. et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* **20**, 249 (2019).
3. Campisi, J. et al. From discoveries in ageing research to therapeutics for healthy ageing. *Nature* **571**, 183–192 (2019).
4. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. Hallmarks of aging: an expanding universe. *Cell* **186**, 243–278 (2023).
5. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
6. Deng, Y. T. et al. Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell* **188**, 253–271.e257 (2025).
7. de Magalhães, J. P. Cellular senescence in normal physiology. *Science* **384**, 1300–1301 (2024).
8. Dormann, D. & Lemke, E. A. Adding intrinsically disordered proteins to biological ageing clocks. *Nat. Cell Biol.* **26**, 851–858 (2024).
9. Oh, H. S. et al. Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).
10. Wang, J. et al. Accurate estimation of biological age and its application in disease prediction using a multimodal image Transformer system. *Proc. Natl Acad. Sci. USA* **121**, e2308812120 (2024).
11. Dörfel, R. P. et al. Prediction of brain age using structural magnetic resonance imaging: a comparison of accuracy and test-retest reliability of publicly available software packages. *Hum. Brain Mapp.* **44**, 6139–6148 (2023).
12. Raghu, V. K., Weiss, J., Hoffmann, U., Aerts, H. & Lu, M. T. Deep learning to estimate biological age from chest radiographs. *JACC Cardiovasc. Imaging* **14**, 2226–2236 (2021).
13. Zhu, Z. et al. Retinal age gap as a predictive biomarker for mortality risk. *Br. J. Ophthalmol.* **107**, 547–554 (2023).
14. Chen, R. et al. Biomarkers of ageing: current state-of-art, challenges and opportunities. *MedComm Future Med.* **2**, e50 (2023).
15. Kivimäki, M. et al. Proteomic organ-specific ageing signatures and 20-year risk of age-related diseases: the Whitehall II observational cohort study. *Lancet Digit. Health* **7**, e195–e204 (2025).
16. Hou, Y. et al. Ageing as a risk factor for neurodegenerative disease. *Nat. Rev. Neurol.* **15**, 565–581 (2019).
17. Bafei, S. E. C. & Shen, C. Biomarkers selection and mathematical modeling in biological age estimation. *npj Aging* **9**, 13 (2023).
18. Yousefzadeh, M. J. et al. An aged immune system drives senescence and ageing of solid organs. *Nature* **594**, 100–105 (2021).
19. Wang, K. et al. Epigenetic regulation of aging: implications for interventions of aging and diseases. *Signal Transduct. Target. Ther.* **7**, 374 (2022).
20. Xia, X., Chen, W., McDermott, J. & Han, J. J. Molecular and phenotypic biomarkers of aging. *F1000Res.* **6**, 860 (2017).
21. Goeminne, L. J. et al. Plasma protein-based organ-specific aging and mortality models unveil diseases as accelerated aging of organismal systems. *Cell Metab.* **37**, 205–222.e206 (2025).
22. Elliott, M. L. et al. Disparities in the pace of biological aging among midlife adults of the same chronological age have implications for future frailty risk and policy. *Nat. Aging* **1**, 295–308 (2021).
23. Jylhava, J., Pedersen, N. L. & Hagg, S. Biological age predictors. *EBioMedicine* **21**, 29–36 (2017).
24. Lu, A. T. et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303–327 (2019).
25. Foy, B. H. et al. Haematological setpoints are a stable and patient-specific deep phenotype. *Nature* **637**, 430–438 (2025).
26. Alberti, S. & Hyman, A. A. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nat. Rev. Mol. Cell Biol.* **22**, 196–213 (2021).
27. Tang, A. S. et al. Harnessing EHR data for health research. *Nat. Med.* **30**, 1847–1855 (2024).
28. Heumos, L. et al. An open-source framework for end-to-end analysis of electronic health record data. *Nat. Med.* **30**, 3369–3380 (2024).
29. Mataraso, S. J. et al. A machine learning approach to leveraging electronic health records for enhanced omics analysis. *Nat. Mach. Intell.* **7**, 293–306 (2025).
30. Zhang, K. et al. Concepts and applications of digital twins in healthcare and medicine. *Patterns (N. Y.)* **5**, 101028 (2024).
31. Rutledge, J., Oh, H. & Wyss-Coray, T. Measuring biological age using omics data. *Nat. Rev. Genet.* **23**, 715–727 (2022).
32. Deng, Y. Digital twin-based modeling of complex systems for smart aging. *Discret. Dyn. Nat. Soc.* **2022**, 7365223 (2022).
33. Thompson, D. J. et al. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. Preprint at https://www.medrxiv.org/content/10.1101/2022.06.16.22276246v1 (2022).
34. Cao, Z. J. & Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
35. Tuttle, C. S. L., Thang, L. A. N. & Maier, A. B. Markers of inflammation and their association with muscle strength and mass: a systematic review and meta-analysis. *Ageing Res. Rev.* **64**, 101185 (2020).
36. Ndumele, C. E. et al. A synopsis of the evidence for the science and clinical management of Cardiovascular-Kidney-Metabolic (CKM) Syndrome: a scientific statement from the American Heart Association. *Circulation* **148**, 1636–1664 (2023).
37. Ronan, V., Yeasin, R. & Claud, E. C. Childhood development and the microbiome—the intestinal microbiota in maintenance of health and development of disease during childhood development. *Gastroenterology* **160**, 495–506 (2021).
38. Chen, Q. et al. OMICmAge: an integrative multi-omics approach to quantify biological age with electronic medical records. Preprint at https://www.biorxiv.org/content/10.1101/2023.10.16.562114v1 (2023).
39. Garg, M. et al. Disease prediction with multi-omics and biomarkers empowers case-control genetic discoveries in the UK Biobank. *Nat. Genet.* **56**, 1821–1831 (2024).
40. Sahraeian, S. M. E. et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat. Commun.* **8**, 59 (2017).

41. Shuken, S. R. et al. Limited proteolysis-mass spectrometry reveals aging-associated changes in cerebrospinal fluid protein abundances and structures. *Nat. Aging* **2**, 379–388 (2022).
42. Tian, Y. E. et al. Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nat. Med.* **29**, 1221–1231 (2023).
43. de Magalhães, J. P. Distinguishing between driver and passenger mechanisms of aging. *Nat. Genet.* **56**, 204–211 (2024).
44. de Magalhães, J. P. et al. Human Ageing Genomic Resources: updates on key databases in ageing research. *Nucleic Acids Res* **52**, D900–D908 (2024).
45. Pereira, J. B. et al. DOPA decarboxylase is an emerging biomarker for Parkinsonian disorders including preclinical Lewy body disease. *Nat. Aging* **3**, 1201–1209 (2023).
46. Qiu, W., Chen, H., Kaeberlein, M. & Lee, S. I. ExplaiNAble BioLogical Age (ENABL Age): an artificial intelligence framework for interpretable biological age. *Lancet Healthy Longev.* **4**, e711–e723 (2023).
47. Sun, T., He, X., Song, X., Shu, L. & Li, Z. The digital twin in medicine: a key to the future of healthcare?. *Front. Med. (Lausanne)* **9**, 907066 (2022).
48. Moqri, M. et al. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* **186**, 3758–3775 (2023).
49. Sun, E. D. et al. Spatial transcriptomic clocks reveal cell proximity effects in brain ageing. *Nature* **638**, 160–171 (2025).
50. Unger Avila, P. et al. Gene regulatory networks in disease and ageing. *Nat. Rev. Nephrol.* **20**, 616–633 (2024).
51. Wang, T. W. et al. Blocking PD-L1-PD-1 improves senescence surveillance and ageing phenotypes. *Nature* **611**, 358–364 (2022).
52. Di Micco, R., Krizhanovsky, V., Baker, D. & d'Adda di Fagagna, F. Cellular senescence in ageing: from mechanisms to therapeutic opportunities. *Nat. Rev. Mol. Cell Biol.* **22**, 75–95 (2021).
53. Gorbunova, V. et al. The role of retrotransposable elements in ageing and age-associated diseases. *Nature* **596**, 43–53 (2021).
54. Leote, A. C., Lopes, F. & Beyer, A. Loss of coordination between basic cellular processes in human aging. *Nat. Aging* **4**, 1432–1445 (2024).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Kai Wang [1,2,3,23], Fei Liu [4,5,23], Wei Wu [2,3,23], Changxi Hu [2,23], Xian Shen [1,23], Meihao Wang [6,7,23], Gen Li [2,23], Fanxin Zeng [8,23], Li Liu [9,23], Io Nam Wong [4], Sian Liu [2], Zixing Zou [10], Bingzhou Li [2,10], Jinghang Li [11], Xiaoying Huang [12], Shengwei Jin [13], Zhuomin Li [2], Hui Xu [2], Gang Chen [1], Xiaodong Chen [1], Ying Zhu [6,7], Ping Li [14], Zhe Feng [14], Winston Wang [15], Linling Cheng [4], Mingqi Yang [4,16], Qiang Hou [2], Wenyang Lu [2], Yiwen Sun [17], Kun Li [2], Tian Zhong [4], Zhuo Sun [2,18], Yun Yin [2,19], Alexandre Loupy [20], Eric Oermann [21], Xiangmei Chen [14] ✉, Kang Zhang [2,4,10] ✉ & for the International Consortium of Digital Twins in Healthcare and Medicine*

[1]Department of General Surgery, Department of Hepatobiliary Surgery, Zhejiang Key Laboratory of Intelligent Cancer Biomarker Discovery and Translation, Zhejiang-Germany Interdisciplinary Joint Laboratory of Hepatobiliary-Pancreatic Tumor and Bioengineering, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China. [2]State Key Laboratory of Eye Health, Eye Hospital, Clinical Data Science Institute, Institute for Advanced Study on Eye Health and Diseases, Wenzhou Medical University, Wenzhou, China. [3]Department of Big Data and Biomedical AI, College of Future Technology, Peking University and Peking-Tsinghua Center for Life Sciences, Beijing, China. [4]Institute for AI in Medicine and Faculty of Medicine, Macau University of Science and Technology, Macau, China. [5]National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College Beijing, Beijing, China. [6]Department of Radiology, The Second Affiliated Hospital of Wenzhou Medical University, Wenzhou, China. [7]Key Laboratory of Intelligent Medical Imaging of Wenzhou, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China. [8]Department of Clinical Research Center, Sichuan Province Clinical Medical Research Center for Imaging Medicine, Dazhou Central Hospital, Dazhou, Sichuan, China. [9]Department of Health Management and Department of Infectious Diseases, Nanfang Hospital, Southern Medical University, Guangzhou, China. [10]Guangzhou National Laboratory, Guangzhou, China. [11]University of Pittsburgh, Department of Bioengineering, Pittsburgh, PA, USA. [12]Division of Pulmonary Medicine, the First Affiliated Hospital, Wenzhou Medical University, Wenzhou Key Laboratory of Interdisciplinary and Translational Medicine, Wenzhou Key Laboratory of Heart and Lung, Wenzhou, Zhejiang, China. [13]Department of Anesthesia and Critical Care, The Second Affiliated Hospital and Yuying Children's Hospital of Wenzhou Medical University,Key Laboratory of Pediatric Anesthesiology, Ministry of Education, Wenzhou Medical University, Wenzhou, Zhejiang, China. [14]Department of Nephrology, First Medical Center of Chinese PLA General Hospital, State Key Laboratory of Kidney Diseases, National Clinical Research Center of Kidney Diseases, Beijing Key Laboratory of Kidney Disease, Beijing, China. [15]Mayo Clinic Department of Internal Medicine, Scottsdale, AZ, USA. [16]Zhuhai People's Hospital, The Affiliated Hospital of Beijing Institute of Technology, Zhuhai Clinical Medical College of Jinan University, Zhuhai, Guangdong, China. [17]Nepean Hospital, Sydney, Australia. [18]Department of Ophthalmology, The Third People's Hospital of Changzhou, Changzhou, China. [19]Faculty of Health and Wellness, Faculty of Business, City University of Macau, Macau, SAR, China. [20]Université Paris Cité, INSERM U970 PARCC, Paris Institute for Transplantation and Organ Regeneration, Paris, France. [21]Departments of Neurosurgery, Radiology, and Data Sceince, Neuroscience Institute, NYU Langone Medical Center, New York University, New York, NY, USA. [23]These authors contributed equally: Kai Wang, Fei Liu, Wei Wu, Changxi Hu, Xian Shen, Meihao Wang, Gen Li, Fanxin Zeng, Li Liu. * A list of authors and their affiliations appears at the end of the paper. ✉e-mail: xmchen301@126.com; kang.zhang@gmail.com

for the International Consortium of Digital Twins in Healthcare and Medicine

Kai Wang[1,2,3,23], Fei Liu[4,5,23], Wei Wu[2,3,23], Xian Shen[1,23], Meihao Wang[6,7,23], Fanxin Zeng[8,23], Li Liu[9,23], Io Nam Wong[4], Manson Fok[4], Taiwa Hou[22], Jinghang Li[11], Xiaoying Huang[12], Shengwei Jin[13], Lei Guo[13], Miaosang Xu[13], Dan Yao[12], Chengjing Wang[12], Pingzhen Yang[16], Caiwen Ou[9], Xueqiang Wang[13], Aimin Wu[13], Gang Chen[1], Xiaodong Chen[14], Winston Wang[16], Yiwen Sun[17], Alexandre Loupy[20], Eric Oermann[21], Xiangmei Chen[15] & Kang Zhang[2,4,10]

[22]Conde S. Januário Hospital, Macau, China.

Methods

Study populations

The China Health Aging Investigation (CHAI), as a project of the International Consortium of Digital Twin in Medicine[30], is an ongoing study using EHRs to predict patients' BA and assess individual disease risks[10,55–57]. Data for this study were sourced from several hospitals in the CHAI project. Cohort #1 (The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China), cohort #2 (The Second Affiliated Hospital of Wenzhou Medical University, Wenzhou, China), cohort #4 (Dazhou People Hospital, Sichuan, China) and cohort #5 (Nanfang Hospital, Southern Medical University, Guangzhou, China and the PLA General Hospital, Beijing, China) are major tertiary hospitals offering full comprehensive adult services, whereas cohort #3 (Women and Children's Center of the PLA General Hospital and Women and Children's Center of the Second Affiliated Hospital of Wenzhou Medical University, China) comprises major regional referral hospitals with primary services focused on women and children's health and diseases. Our analysis included 24,633,025 longitudinal clinical visits from the EHR data of 9,680,764 patients. Additionally, longitudinal EHR data from cohort #5 and the UK Biobank were utilized as two external validation cohorts. Data were collected on biological sex. Ethics Committee approvals were obtained in all institutions. The study was registered at clinicaltrial.gov (NCT06791486). The work was conducted in compliance with the Chinese CDC policy on reportable infectious diseases and the Chinese Health and Quarantine Law, in compliance with patient privacy regulations in China, and was adherent to the tenets of the Declaration of Helsinki. For the purposes of training our biological clock, 'healthy' individuals were defined as participants who had no recorded disease diagnoses within their EHRs at the time of their clinical visits. This approach was important for establishing a baseline model of a normal pediatric development clock and an adult aging clock, against which BA deviations in individuals with specific diseases could be precisely assessed.

Data representation

We structured the EHR data as chronological sequences of clinical visits for each patient. Each patient's longitudinal clinical record is represented as a time-ordered sequence $S = \{(X_0, T_0), (X_1, T_1), \ldots, (X_L, T_L)\}$, where X_i denotes the vector of clinical variables (including continuous and categorical laboratory test results and clinical measurements) collected at the ith visit, T_i represents the time elapsed (in days) since the initial visit, with $T_0 = 0$ by definition, and L is the number of visits for this patient. The continuous clinical variables were quantized according to the formula $D(x) = \lfloor (x - X_{max})/(X_{max} - X_{min}) \times d_{\text{cont}} \rfloor$, where $\lfloor X \rfloor$ represents the floor function, X_{max} the maximum value of feature x, X_{min} the minimum value of feature x, and d_{cont} is the number of discrete bins. This discretization resulted in integer values between 0 and d_{cont}, with values exceeding the defined range truncated to the maximum boundary and missing values encoded as −1. This discretization strategy preserved the distributional characteristics of the original variables while enabling a unified representation of patient data. At each visit, features X_i were represented as a concatenation of categorical variables and discretized continuous variables: $X_i = [X_{\text{cat}}; X_{\text{cont}}]$, where $X_{\text{cat}} \in \mathbb{N}^{L \times N_{\text{cat}}}$ and $X_{\text{cont}} \in \mathbb{N}^{L \times N_{\text{cont}}}$, respectively, with L denoting the number of clinical visits, N_{cat} and N_{cont} are the numbers of categorical and continuous features, respectively.

EHRFormer architecture

EHRFormer is an encoder–decoder style transformer architecture specifically designed to process longitudinal EHR data. The model comprises three key components: an examination encoder, a temporal embedding and task-specific decoder heads.

EHRFormer architecture and examination encoder

After preprocessing each patient's longitudinal EHR data through discretization and concatenation into a unified feature representation as $X_i = [X_{\text{cat}}; X_{\text{cont}}]$, we implemented a visit-level encoding framework. Similar to a BERT's[58] embedding approach, our embedding layer employed a dual representation strategy: discretized feature values were encoded using shared token embeddings to represent their magnitude, and separate type embeddings were assigned to each variable position to denote the specific clinical feature category. This complementary embedding method allowed the model to simultaneously capture both the value distributions and the semantic meaning of different clinical measurements. A designated special vector reserved (preserved) missing examinations, enabling the model to differentiate between absent tests and actual clinical observations. To capture complex interdependencies between clinical variables, we applied a transformer-based architecture that processed these embedded features through multiple self-attention layers. This encoding process can be formalized as $E_{\text{visit}} = \text{Encoder}(\text{Embed}(X_i))$, where Encoder is a Transformer encoder that generates a contextualized representation for each clinical visit.

EHRFormer architecture, temporal embedding and decoder

To model disease progression and capture the longitudinal nature of patient trajectories, we implemented a temporal embedding to capture the relative time between visits. From the examination encoder output, we retrieved a visit-level embedding, E_{visit}. To model temporal relationships, we used days elapsed since the initial visit as a linear positional embedding TimeEmbed (T) to enable the architecture to learn time-dependent patterns in longitudinal EHR data. To create a longitudinal patient-level representation, we passed visit embeddings E_{visit} augmented with time information through a Transformer decoder: $E_{\text{patient}} = \text{Decoder}\,(E_{\text{visit}} + \text{TimeEmbed}\,(T))$, where causal masking ensures unidirectional information flow in this autoregressive process.

EHRFormer architecture and task-specific decoders

Following the established patient-level longitudinal representation E_{patient}, we designed a task-specific decoder with separate pathways for discrete outcomes (for example, diagnosis prediction) and continuous measurements (for example, biomarker estimation and BA prediction). Each pathway applies a projection layer followed by ReLU activation, formalized as $y_i = \text{ReLU}(W_i^{\text{T}} E_{\text{patient},\,i})$, where $E_{\text{patient},\,i}$ represents a patient's digital representation derived from first to ith visit. Importantly, causal masking prevents information from future visits from influencing predictions at the ith visit, ensuring fairness by restricting the model to only information available in real clinical scenarios. This architecture also enables simultaneous handling of diverse clinical prediction tasks while facilitating knowledge transfer between related objectives through jointly optimized parameters.

Training procedures

Our training procedure consisted of two stages: pretraining and fine-tuning. During pretraining, we employed self-supervised learning on unlabeled longitudinal EHR data to develop robust clinical representations. The subsequent fine-tuning stage adapted these representations for specific prediction tasks. This approach leverages generalizable patterns from large-scale unlabeled data before specializing downstream applications. Both stages utilize specialized loss functions and incorporate strategies to mitigate dataset-specific biases.

Controlling for missingness and cohort bias through adversarial methods

Missing values in EHRs lead to incomplete or biased digital representations, as models may inadvertently learn to rely on the missing-state biases rather than the true clinical meaning of the feature expression values. Drawing inspiration from examples of the concept of adversarial learning in other domains, we implemented a missingness discriminator output head. Concurrently, the missingness discriminator is designed to determine whether a specific feature value is missing or not. We implemented a gradient reversal layer (GRL) between the feature encoder and the missingness discriminator. During backpropagation, the GRL inverts the gradient, compelling the feature encoder to produce representations that are independent of the missingness status. This forces the encoder to focus on encoding the inherent clinical importance of the features rather than being influenced by whether a value is present or absent. The loss function for the missingness discrimination task is defined as $\mathcal{L}_{\text{missing}} = -\frac{1}{M}\sum_{j=1}^{M}\sum_{m=0}^{1} z_{j,m}\log\hat{z}_{j,m}$, where $z_{j,m}$ is a binary indicator denoting whether the feature value of sample j has a missing status m (0 for present, 1 for missing), $\hat{z}_{j,m}$ is the predicted probability, and M is the number of samples with potential missing values. By minimizing this loss, the model is encouraged to learn missing-invariant representations. This approach enables the creation of more robust digital representations that can better generalize across datasets with different missing value patterns, ultimately improving the accuracy and reliability of clinical outcome predictions.

Cohort bias, also known as a batch effect, is a substantial challenge in multi-center data studies. Clinical data collected from different hospitals often exhibit systematic variations due to differences in patient demographics, practice patterns and measurement protocols, potentially leading to biased models. Similar to the missingness discriminator, we designed a cohort discriminator that aims to identify the cohort label of each sample, while the encoder is forced to suppress cohort-specific information. The cohort classification loss is formulated as $\mathcal{L}_{\text{cohort}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} y_{i,d}\log(\hat{y}_{i,d})$, where $y_{i,d}$ is a binary indicator of whether sample i belongs to domain d, $\hat{y}_{i,d}$ is the predicted probability, N is the number of samples, and D is the number of domains (clinical cohorts). This approach encourages the model to learn cohort-invariant representations that generalize across healthcare settings while maintaining predictive performance for clinical outcomes.

Pretraining step

We employed a self-supervised pretraining approach with multiple complementary objectives to enable our model to learn comprehensive representations of EHR data. We randomly masked 50% of the valid test results in the current examination as input, and trained the model to predict 50% masked values in the current examination and next examination. The masked language modeling loss function is defined as $\mathcal{L}_{\text{MLM}} = \frac{1}{|\mathcal{M}|}\sum_{i=0}^{N}\sum_{j\in\mathcal{M}_i}\mathcal{L}_{\text{MSE}}(\hat{v}_{i,j}, v_{i,j})$, where $\mathcal{M}_i$ is the set of masked indices in examination event s_i and the next examination after s_i, $|\mathcal{M}|$ is the total number of masked tokens across all examination events, $v_{i,j}$ is the true value of the jth test in examination s_i and the next examination after s_i, and $\hat{v}_{i,j}$ is the predicted value. To quantify uncertainty in the clinical data, we incorporated a variational framework with evidence lower bound (ELBO) maximization as $\mathcal{L}_{\text{ELBO}} = E_{q_\phi}[\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x)||p_\theta(z|x))$, balancing reconstruction fidelity against latent space regularization. Additionally, we incorporated the domain adversarial loss $\mathcal{L}_{\text{domain}}$ and $\mathcal{L}_{\text{missing}}$ to promote cohort-invariant and missing-invariant representations. Finally, for the age regression task, we trained the model to predict patients' ages at each examination event using only clinical measurements (with all age-related information explicitly removed from inputs) to assess biological aging patterns, using the mean squared error (MSE) loss function defined as $\mathcal{L}_{\text{age}} = \frac{1}{N}\sum_{i=0}^{N}(\hat{a}_i - a_i)^2$, where $\hat{a}_i$ is the predicted age at the examination event s_i, and a_i is the true age. Only healthy individuals were included in the loss calculation for the age-prediction task, enabling EHRFormer to construct a biological clock reflecting normal aging patterns. This approach allows subsequent precise assessment of BA deviations between diseased individuals and their healthy peers in subsequent CA–BA differential analyses. Therefore, the final pretraining objective combined these components with appropriate weighting coefficients: $\mathcal{L}_{\text{pretrain}} = \alpha_1\mathcal{L}_{\text{MLM}} + \alpha_2\mathcal{L}_{\text{ELBO}} - \alpha_3\mathcal{L}_{\text{cohort}} - \alpha_4\mathcal{L}_{\text{missing}} + \alpha_5\mathcal{L}_{\text{age}}$, where the negative sign reflects the gradient reversal mechanism.

Fine-tuning step for disease state prediction tasks

We implemented three distinct disease prediction tasks that reflect different clinical scenarios: first occurrence disease diagnosis, future disease prediction and fixed-time-window future prediction.

For first occurrence disease diagnosis, we trained the model to identify the first occurrence of specific diseases, excluding subsequent visits after initial diagnosis to capture true onset patterns rather than disease management. Formally, for a patient with a longitudinal sequence S with length L, and where $l_{i,d}$ represents whether this patient was diagnosed as positive for disease d at the ith visit, the label $c_{i,d}$ of the first occurrence diagnosis task is defined as

$$c_{i,d} = \begin{cases} 1, \text{ if } l_{i,d} = 1 \text{ and } l_{j,d} = 0 \text{ for all } j < i \\ 0, \text{ if } l_{i,d} = 0 \text{ and } l_{j,d} = 0 \text{ for all } j \in \{0, 1, \ldots, L\}. \end{cases}$$

For the future disease prediction task, we developed a labeling strategy to identify patients at risk before disease manifestation, using each visit as a dynamic baseline for prediction. Formally, for a patient with longitudinal sequence S with length L, the label $f_{i,d}$ of the future prediction task for disease d at the ith visit is defined as

$$f_{i,d} = \begin{cases} 1, \text{ if } l_{j,d} = 0 \text{ for } j \leq i \text{ and } \exists\, k > i \text{ such that } l_{k,d} = 1 \\ 0, \text{ if } l_{i,d} = 0 \text{ and } l_{j,d} = 0 \text{ for all } j \in \{0, 1, \ldots, L\}. \end{cases}$$

The third prediction task assesses N-year disease incidence. This is achieved by predicting over a fixed look-ahead window ($t = 5$ or 10 years) from each potential per-visit baseline. To ensure the validity of our labels, we implemented rigorous censoring for any observation with insufficient follow-up time. Formally, for a patient with recorded age $A(i)$, the rolling t-year window prediction label $w_{i,d}^{t}$ of disease d at visit i is defined as

$$w_{i,d}^{t} = \begin{cases} 1, \text{ if } l_{j,d} = 0 \text{ for } j \leq i \text{ and } \exists\, k > i \text{ such that } l_{k,d} = 1 \text{ and } A(k) - A(i) \leq t \\ 0, \text{ if } l_{i,d} = 0 \text{ and } l_{j,d} = 0 \text{ for all } j \in \{0, 1, \ldots, L\} \text{ and } A(L) - A(i) \geq t. \end{cases}$$

The loss function for each task is $\mathcal{L} = \frac{1}{N}\sum_{i=0}^{N}\sum_{d=1}^{D}\mathcal{L}_{\text{BCE}}(\hat{y}_{i,d}, y_{i,d})$, where $\hat{y}_{i,d}$ is the predicted probability of disease d on one of the above three labels, D is the total number of diseases considered, and $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy loss.

Implementation details

We implemented our EHRFormer architecture using a combination of transformer models. Specifically, we utilized a 24-layer transformer encoder with a hidden dimension of 1,024 as the examination encoder to process individual examination events, and a 12-layer autoregressive transformer decoder with a hidden dimension of 768 as the temporal encoder to capture longitudinal patterns across the sequence of examinations. This design leverages the attention capabilities of the multi-headed self-attention mechanism for understanding relationships between clinical measurements within each examination, while employing the causal masked attention mechanism to model the temporal progression of patient health.

The model was implemented using PyTorch and trained using a two-stage approach. For the pretraining phase, we trained the model for 200 epochs using the Adam optimizer with a learning rate of 10^{-3} and a weight decay of 10^{-6}. The subsequent fine-tuning phase for the downstream tasks was conducted for 100 epochs using the Adam optimizer with a reduced learning rate of 10^{-4}, while maintaining the same weight decay of 10^{-6}.

For both pretraining and fine-tuning steps, we utilized subsets (CHAI-Training and CHAI-Tuning) from the CHAI-Main dataset. Internal validation results were reported using CHAI-Internal, and external validation was conducted using two independent cohorts: CHAI-External-1 and UKB-External. To ensure methodological rigor, we implemented a patient-level non-overlapping partitioning strategy, randomly dividing the CHAI-Main dataset in an 8:1:1 ratio to generate the CHAI-Training, CHAI-Tuning and CHAI-Internal subsets, respectively. The healthy participants in CHAI-Main constituted the CHAI-Healthy Controls cohort used for BA calculation and age difference analysis. The UKB-External dataset comprised all available samples from the UK Biobank cohort.

Age difference calculation

To quantify biological aging deviations, we calculated standardized age differences for each individual using our aging model. First, we predicted BA A_b using the pretrained EHRFormer model on healthy participants in CHAI-Healthy Controls. We then modeled the nonlinear relationship between predicted BA A_b and CA A_c using locally weighted scatterplot smoothing (LOWESS) with a bandwidth parameter of 2/3 via the statsmodels Python package (version 0.14.4) using EHR data from healthy individuals. The resulting function $f(A_c)$ represents the expected BA for a given CA based on healthy population trends. For each individual i, we calculated the raw age difference as $\Delta_i = A_{b,i} - f(A_{c,i})$, representing a deviation from healthy peers with the same CA. Finally, we computed standardized age differences as $z_i = \Delta_i/\sigma$, where σ represents the s.d. of raw age differences within the model.

Visualization of latent space and disease risk analysis

Visualization and clustering of EHRFormer-derived latent vectors were performed by first extracting the laboratory and vital sign features, followed by PCA with 50 components. The resulting embeddings were processed using a neighbor graph approach (15 neighbors, Euclidean metric) and visualized with UMAP (parameters: min_dist=0.3, spread=1.0, 2 components, spectral initialization). Cluster identification was performed using the Leiden community detection algorithm, revealing distinct patient groups that correspond predominantly to pediatric and adult populations. For disease visualization, prevalence and incidence proportions were calculated per cluster. Prevalence was defined as the proportion of individuals with pre-existing disease at baseline (first hospital encounter). Incidence was calculated as the proportion of initially disease-free individuals who developed the condition during the follow-up period (five years from first admission). Each data point was colored according to its corresponding cluster-specific disease prevalence or incidence proportion, providing a visual representation of disease burden across identified patient subgroups. PCA, UMAP and projection visualizations were constructed using the Scanpy[59] Python package (version 1.10.4).

Disease–cluster associations were quantified using adjusted $\log_2$HRs, calculated for each cluster based on the cluster of each patient at their first clinical visit in reference to the remainder of the study population using Cox proportional hazards models. These models incorporated multivariate adjustment for patient demographics (age and sex), smoking, alcohol history and hospital to minimize potential confounding. These associations were visualized using a heatmap with $\log_2$HR values truncated at a maximum of 2 to enhance interpretability while preserving meaningful signal contrast. HRs were calculated using the lifelines Python package (version 0.30.0).

Statistical analysis

We evaluated the performance of regression models for continuous value predictions using MAE, R^2 and PCC. Binary classification models were evaluated using receiver operating characteristic (ROC) curves showing sensitivity versus 1−specificity, with the AUC reported along with 95% confidence intervals. AUCs were calculated using the scikit-learn package (version 1.6.1). Cumulative incidence curves for deciles of disease risk score were calculated using KaplanMeierFitter from the lifelines Python package (version 0.30.0). We plotted cumulative events against each visit age on the x axis. Incidence rates for subsequent records after each given visit age are shown on the y axis.

Reporting Summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

Restrictions apply to the availability of datasets, which were used with the permission of the participants for the current study. Data access requests should be addressed to the corresponding authors and forwarded to a data access committee for approval.

Code availability

Python code for conducting the core analyses is available on GitHub and will be public after publication (https://github.com/kaiwang13/EHRFormer).

References

55. Liang, H. et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* **25**, 433–438 (2019).
56. Zhou, H.-Y. et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* **7**, 743–755 (2023).
57. Tomasev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
58. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* Vol. 1 (long and short papers) (eds Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (ACL, 2019).
59. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

Acknowledgements

This research was supported by the National Natural Science Foundation of China (W2431057 to K.Z., 32141005 and 82030025 to X.C., 32100631 and 32470964 to F.L., 82471955), the Macau Science and Technology Development Fund, Macao (0007/2020/AFJ, 0070/2020/A2 and 0003/2021/AKP to K.Z.), the Basic and Applied Basic Research Project of Guangdong Province (2024A1515220042 to M.Y.), Guangzhou National Laboratory (YW-SLJC0201 to K.Z.), sponsored by the Beijing Nova Program (20240484627 to F.L.), Capital's Funds for Health Improvement and Research (2024-4-40215 to F.L.), the China Postdoctoral Science Foundation (2023T160061 to F.L.) and Macao Young Scholars Program (AM2023018 to F.L.), the Key Disciplines of National Administration of Traditional Chinese Medicine (zyyzdxk-2023310 to X.C.), Lingyan Project of Zhejiang Province (no. 2024C02G1753509) and the Key Laboratory of Intelligent Medical Imaging of Wenzhou (no. 2021HZSY0057). This research was supported by Wenzhou Medical University Eye Health and Disease Advanced Institute.

Author contributions

K.Z. and X.C. conceived, designed and supervised the project. Data collection and analysis were performed by K.W., F.L., W.W., G.L., X.S., M.W., C.H., F.Z., I.N.W., L.L., S.L., Z.Z., B.L., J.L., X.H., S.J., Z.L., H.X., G.C., X.C., Y.Z., P.L., Z.F., W.W., L.C., Q.H., W.L., Y.S., K.L., M.Y., T.Z., Z.S., Y.Y., A.L., E.O., X.C. and K.Z. The manuscript was written by K.W., F.L., W.W., G.L., X.S., M.W., C.H., F.Z., X.C. and K.Z. All authors discussed the results and reviewed the manuscript.

Competing interests

K.W. and K.Z. have filed a patent related to this manuscript. The remaining authors declare no competing interests.

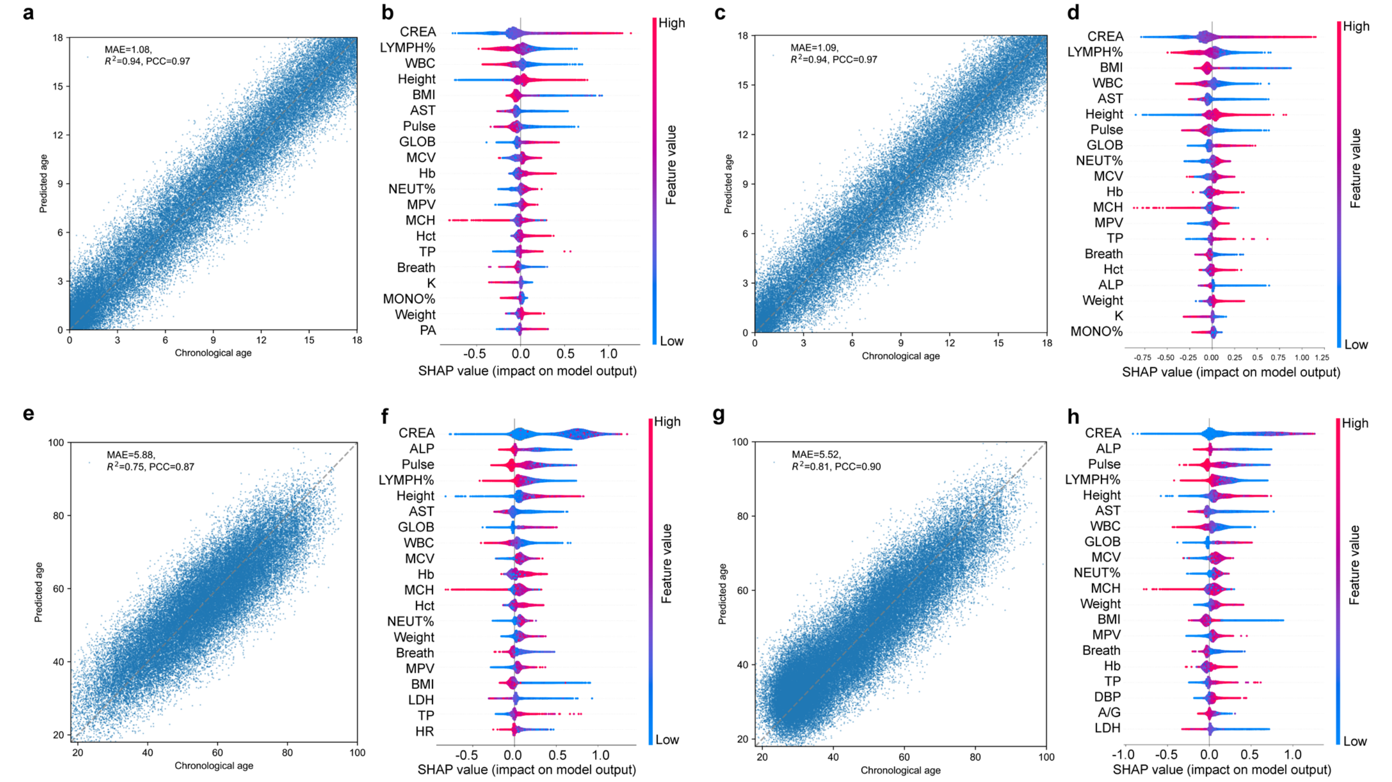
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-025-04006-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-025-04006-w.

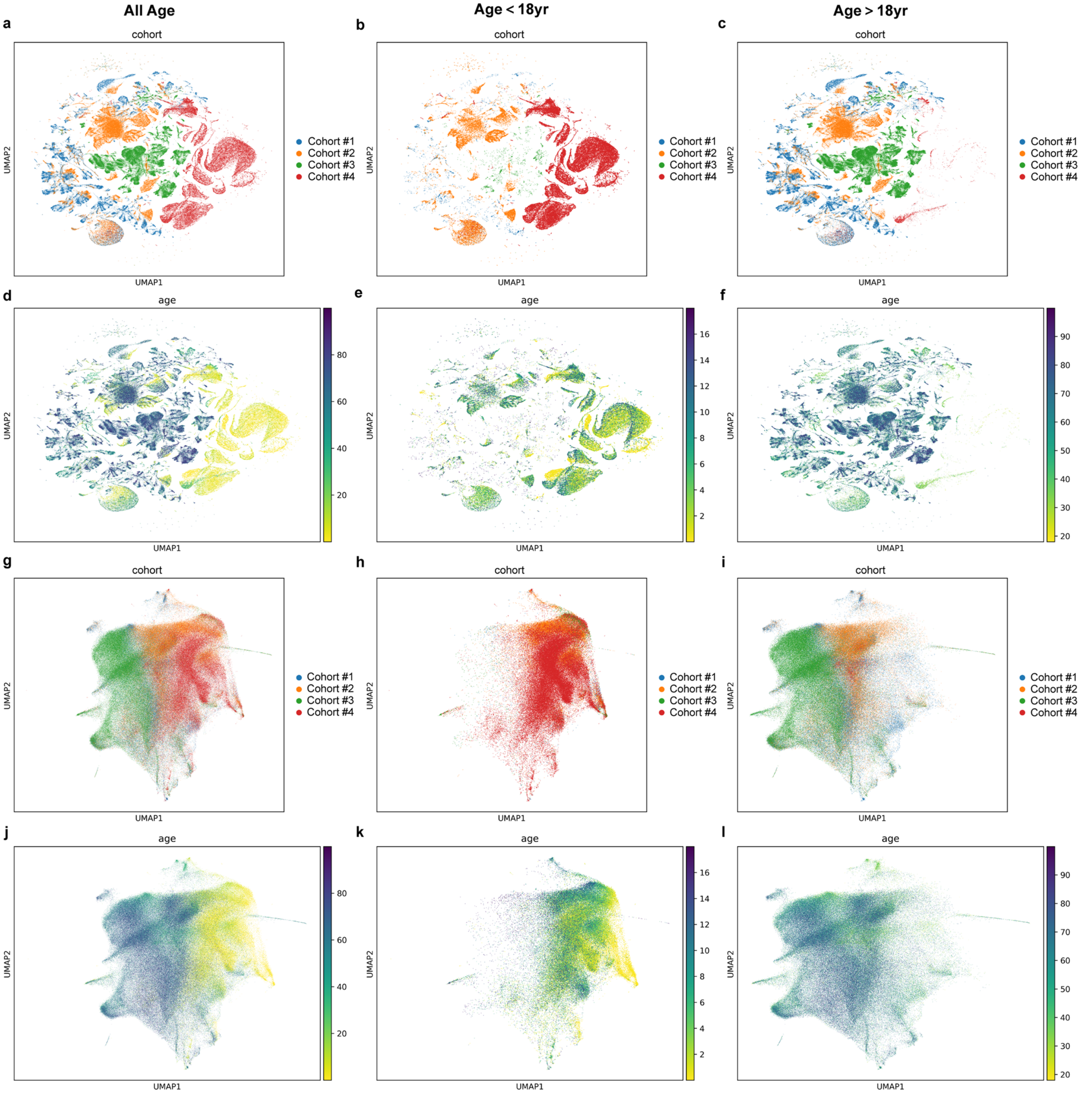
Correspondence and requests for materials should be addressed to Xiangmei Chen or Kang Zhang.

Peer review information *Nature Medicine* thanks M. Austin Argentieri, Nathan Price and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Joao Monteiro, in collaboration with the *Nature Medicine* team.



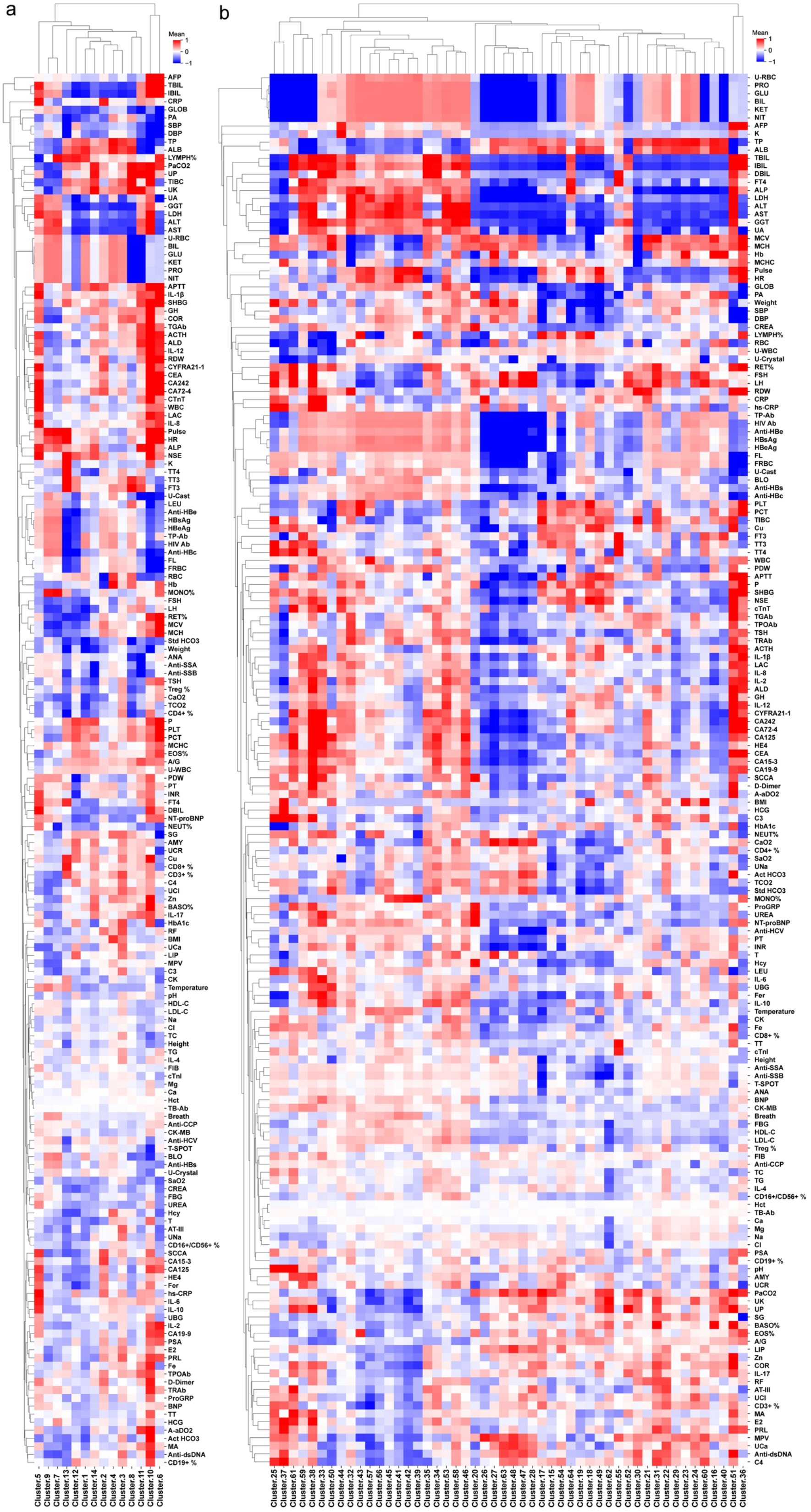
Extended Data Fig. 1 | Aging prediction model in the male and female sexes. **a**. Correlation between actual age and biological age predicted by EHRFormer-based age model, each dot represents one male EHR data point under 18 years old; **b**. The SHAP values of the top 20 contributors in the age prediction for male EHR data under 18 years old; **c**. Correlation between actual age and biological age predicted by EHRFormer-based age model, each dot represents one female EHR data point under 18 years old; **d**. The SHAP values of the top 20 contributors in the age prediction for female EHR data under 18 years old; **e**. Correlation between actual age and biological age predicted by EHRFormer-based age model, each dot represents one male EHR data point over 18 years old; **f**. The SHAP values of the top 20 contributors in the age prediction for male EHR data over 18 years old; **g**. Correlation between actual age and biological age predicted by EHRFormer-based age model, each dot represents one female EHR data point over 18 years old; **h**. The SHAP values of the top 20 contributors in the age prediction for female EHR data over 18 years old.

https://doi.org/10.1038/s41591-025-04006-w



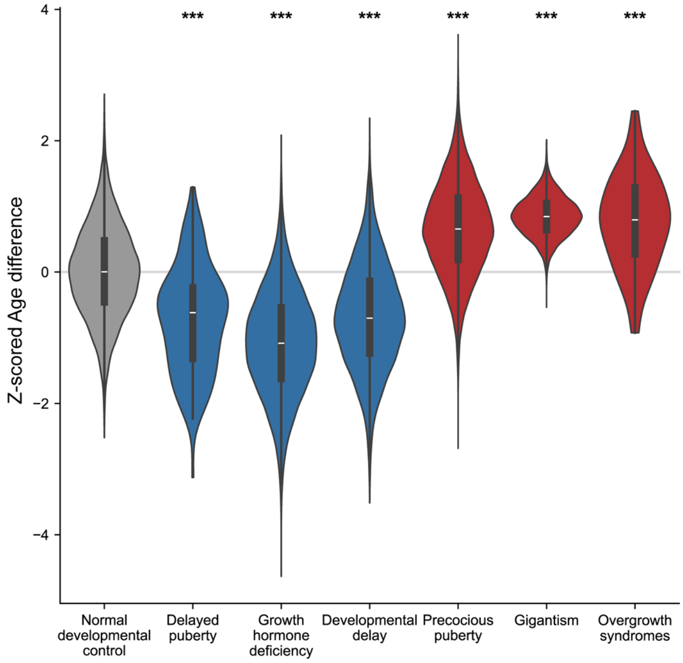
Extended Data Fig. 2 | Elimination of batch effects in hospitals and different datasets. a-c. Cluster analyses showing discrete data clusters of four hospitals before batch effect elimination; **d-f.** Age prediction cluster analyses showing discrete data clusters of four hospitals before batch effect elimination; **g-i.** Cluster analyses showing discrete data clusters of four hospitals after batch effect elimination; **j-l.** Age prediction cluster analyses showing discrete data clusters of four hospitals after batch effect elimination.

https://doi.org/10.1038/s41591-025-04006-w

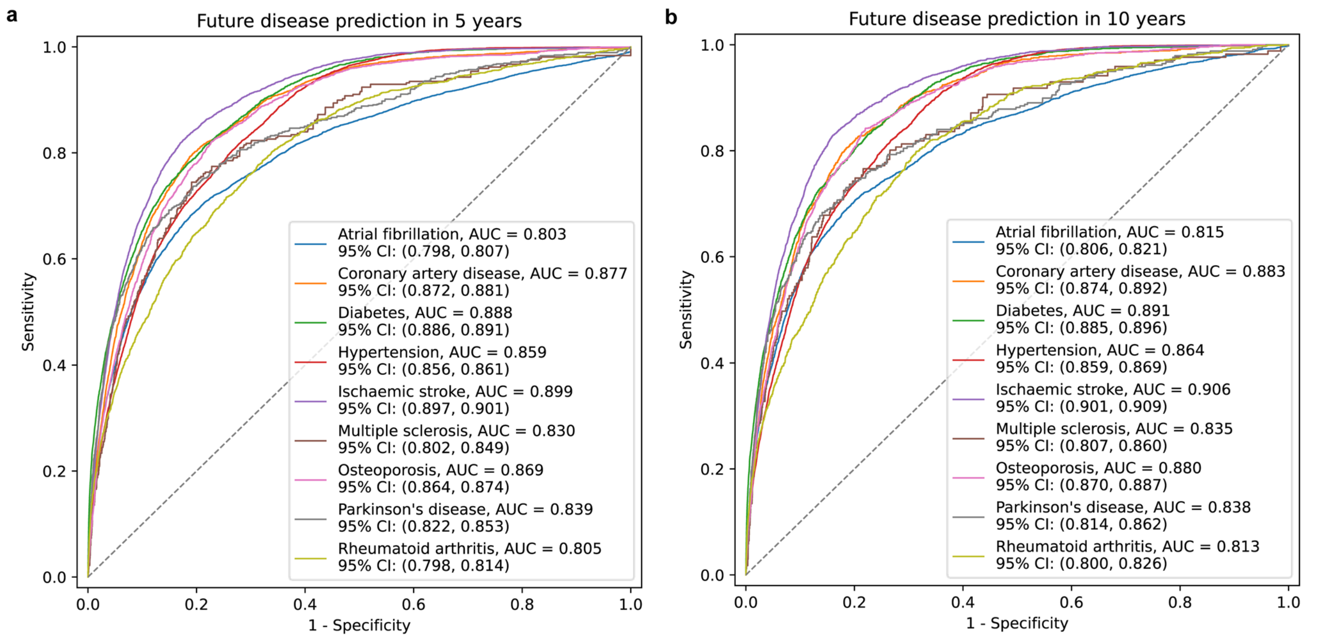


Extended Data Fig. 3 | Correlations of EHR-derived clusters and individual lab test markers. a. pediatric data: y-axis, clusters; x-axis, lab test markers; **b**. adult data: y-axis, clusters; x-axis, lab test markers.

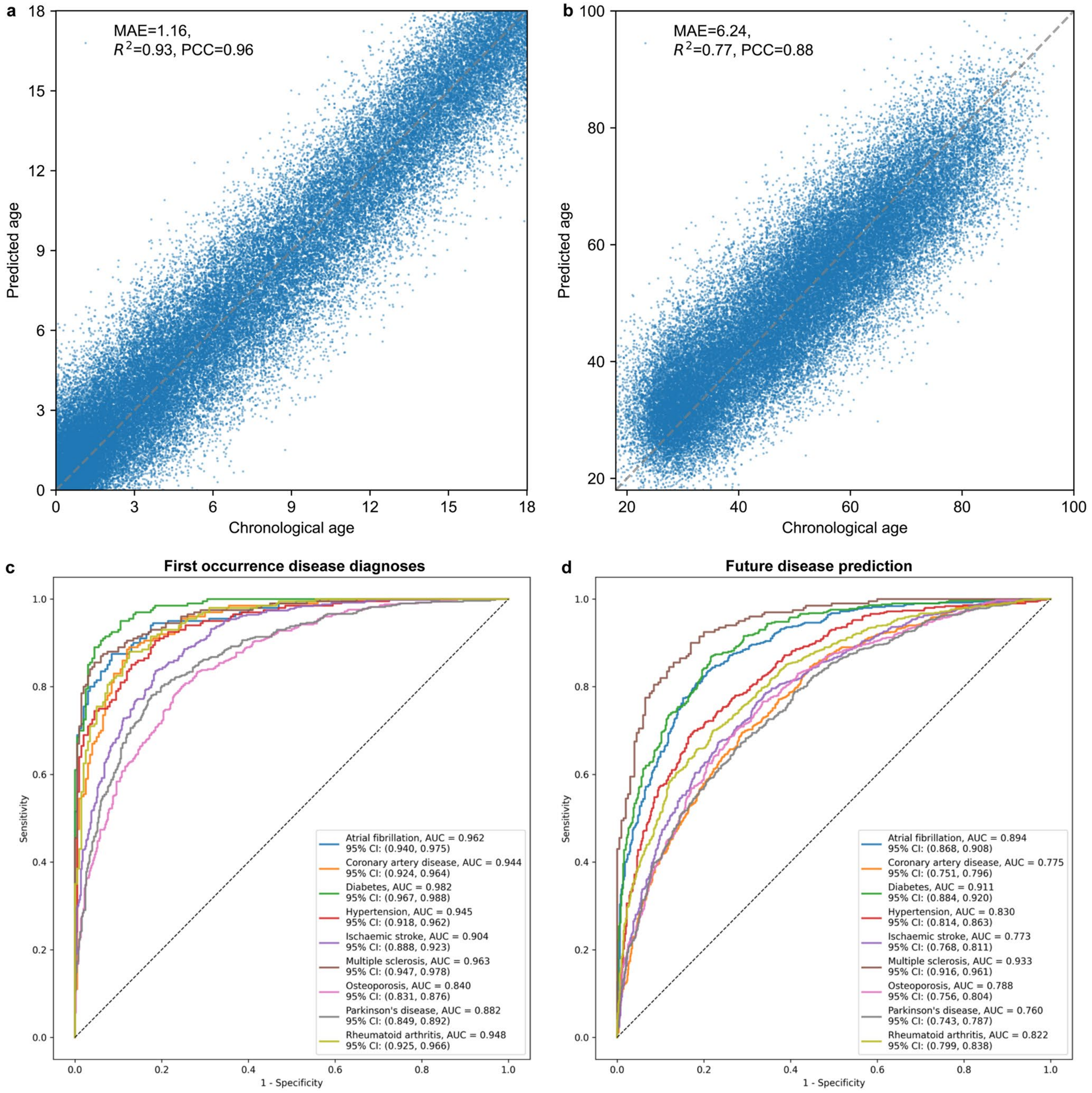
https://doi.org/10.1038/s41591-025-04006-w



Extended Data Fig. 4 | Correlations between selected developmental diseases and developmental clock-derived age differences. Violin graph showing the z-scored age differences of over-maturation group (participants with precocious puberty, gigantism, or overgrowth syndrome) and under-maturation group (participants with delayed puberty, growth hormone deficiency, or developmental delay) compared to normal development control individuals <12. ***: P value < 0.001.

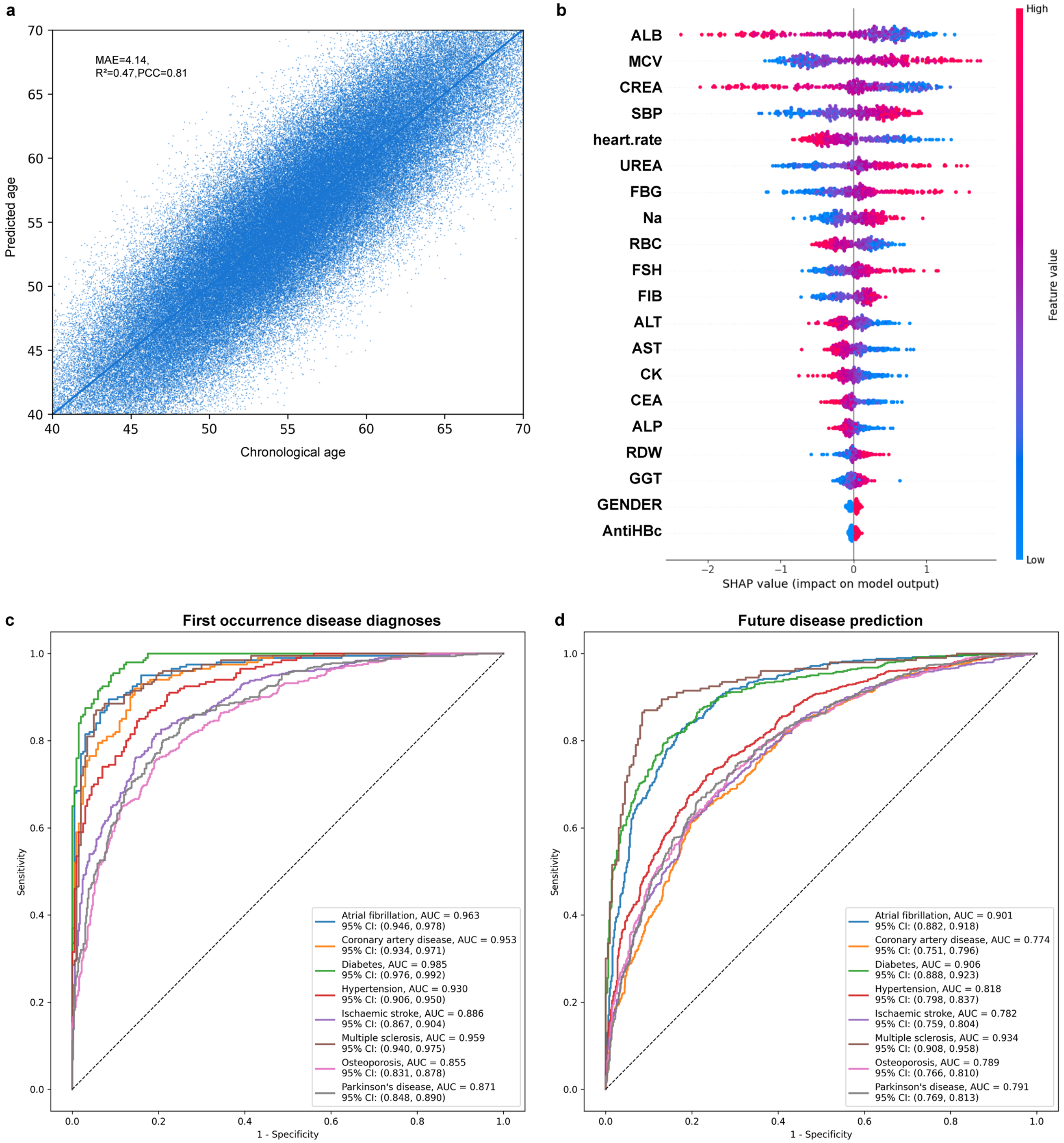


Extended Data Fig. 5 | Validation of EHRFormer-based future disease predictions in 5 and 10 years in CHAI-Internal validation cohort. a. ROC curves of the EHRFormer-based disease prediction model in predicting different diseases in 5 years based on the EHR of internal validation cohort; **b**. ROC curves of the EHRFormer-based disease prediction model in predicting different diseases in 10 years based on the EHR of the internal validation cohort.

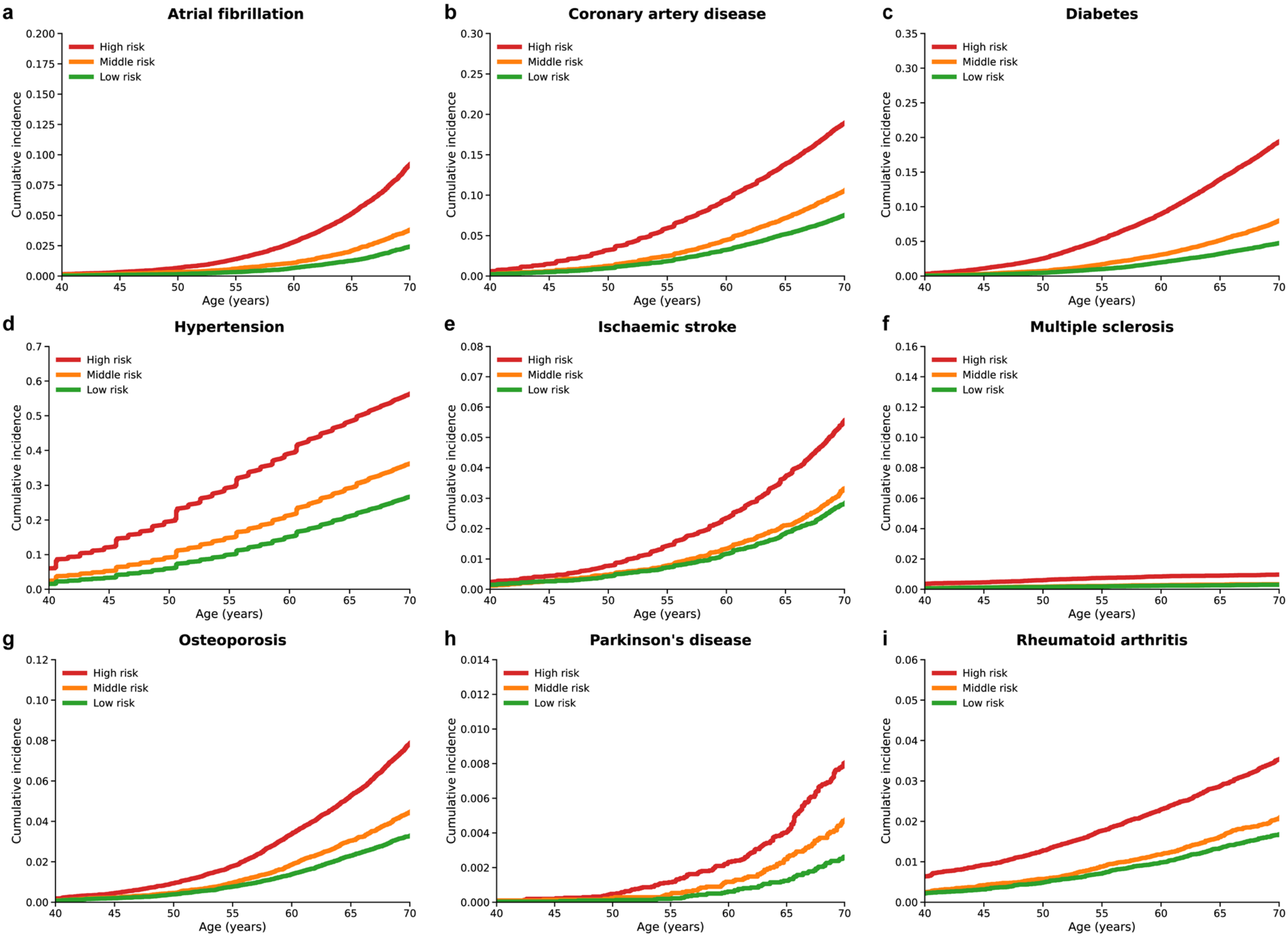


Extended Data Fig. 6 | Validation of EHRFormer-based age and disease prediction models in CHAI-External validation cohort. a. Correlation between CA and BA in the pediatric developmental clock predicted by the EHRFormer-based age model on the EHR of the external validation cohort; **b**. Correlation between CA and BA in the adult aging clock predicted by EHRFormer-based age model on EHR of external validation cohort; **c**. ROC curves of the EHRFormer-based disease prediction model in diagnosing different diseases based on the EHR of the external validation cohort; **d**. ROC curves of the EHRFormer-based disease prediction model in predicting future diseases based on the EHR of the external validation cohort.

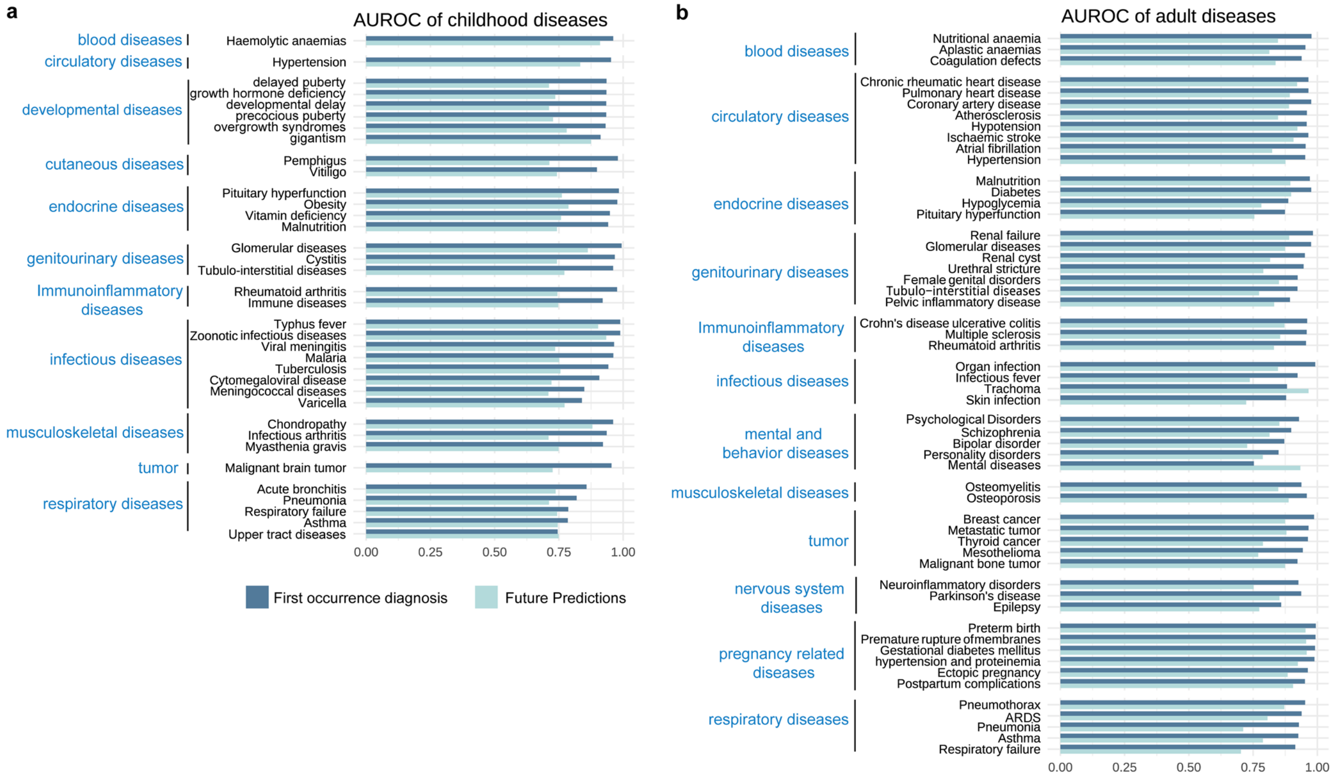
https://doi.org/10.1038/s41591-025-04006-w



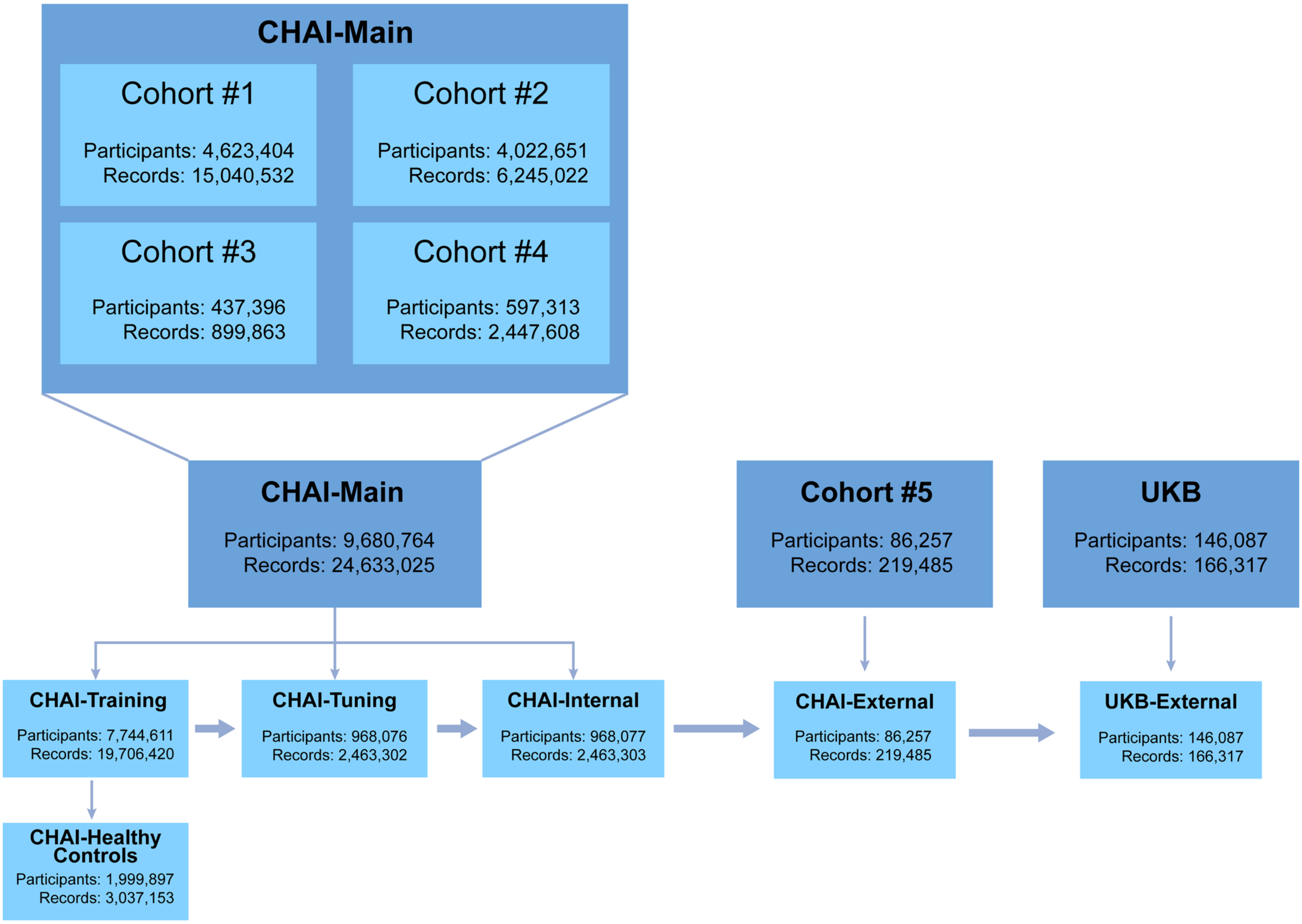
Extended Data Fig. 7 | Validation of EHRFormer-based age and disease prediction models in the UKB-External cohort. a. Correlation between CA and BA in the pediatric developmental clock predicted by an EHRFormer-based age model on the EHR dataset of UKB-External validation cohort; **b**. The SHAP values of the top 20 contributors in the BA prediction for EHR data in UKB-External cohort; **c**. ROC curves of the EHRFormer-based disease prediction model in diagnosing different diseases based on the EHR data of the UKB-external validation cohort; **d**. ROC curves of an EHRFormer-based disease prediction model in predicting future diseases based on EHR data of the UKB-external validation cohort.



Extended Data Fig. 8 | Accumulated risk for various diseases based on a predictive model that categorized individuals into high, middle, and low-risk groups in the UKB validation cohort. a-i. The model demonstrates good predictive capability for all diseases, with distinct separation of risk groups occurring around the age of 40. As time progresses, the gap in accumulated risk between the high, middle, and low-risk groups becomes more pronounced, showcasing the model's ability to predict disease onset and progression over time. Representative diseases include obesity, meningitis, epilepsy, systemic lupus erythematosus, asthma, and arthritis, showcasing the model's ability to predict disease onset and progression.



Extended Data Fig. 9 | Validation of EHRFormer-based disease prediction models in the CHAI-Internal cohort. The performance of the EHRFormer-based disease predicting model in diagnosing and predicting common pediatric and adult diseases using CHAI-Internal <12-year-old **(a)** or >18-year-old **(b)** EHR data.



Extended Data Fig. 10 | Schematic diagrams of cohorts. Four CHAI-training cohorts, CHAI-external validation cohort, and UKB validation cohort.

nature portfolio

Corresponding author(s): Kang Zhang

Last updated by author(s): Aug 29, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No special software or code was used to collect the data.
Data analysis	We used Pytorch for all data analysis. Custom code was at https://github.com/kaiwang13/EHRFormer

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Restrictions apply to the availability of the developmental and validation datasets, which were used with permission of the participants for the current study. Deidentified data may be available for research purposes from the corresponding authors on reasonable request.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We used "sex" in our manuscript because we considered the biological attributes.
Reporting on race, ethnicity, or other socially relevant groupings	Information on race, ethnicity or other socially relevant groupings was not available for all included datasets.
Population characteristics	The China Health Aging Investigation (CHAI), as a project of the International Consortium of Digital Twin in Medicine30, is an ongoing study using EHRs to predict patients' biological age (BA) and assess individual disease risks55-58. Data for this study were sourced from several hospitals in the CHAI project. Cohort #1 (The first affiliated hospital of Wenzhou Medical University, Wenzhou, China), Cohort #2 (The second affiliated hospital of Wenzhou Medical University, Wenzhou, China), Cohort #4 (Dazhou People Hospital, Sichuan, China), and Cohort #5 (Nanfang Hospital, Southern Medical University, Guangzhou, China and the PLA General Hospital, Beijing, China) are major tertiary hospitals offering full comprehensive adult services whereas the cohort #3 (Women and Children's Center of the PLA General Hospital and Women and Children's Center of the second affiliated hospital of Wenzhou Medical University, China) are major regional referral hospitals with primary services focused on women and children's health and diseases. Our analysis included 24,633,025 longitudinal clinical visits from the EHR data of 9,680,764 patients. Additionally, longitudinal EHR data from cohort #5 and the UK Biobank were utilized as two external validation cohorts. Ethics Committee approvals were obtained in all the institutions. The study was registered at clinicaltrial.gov (NCT06791486). The work was conducted in compliance with the Chinese CDC policy on reportable infectious diseases and the Chinese Health and Quarantine Law, and in compliance with patient privacy regulations in China, and was adherent to the tenets of the Declaration of Helsinki. For the purposes of training our biological clock, "healthy" individuals were defined as participants who had no recorded disease diagnoses within their electronic health records (EHRs) at the time of their clinical visits. This approach was important for establishing a baseline model of a normal pediatric development clock and an adult aging clock, against which biological age deviations in individuals with specific diseases could be precisely assessed.
Recruitment	All participants provided written informed consent before enrollment.
Ethics oversight	The study was performed in compliance with the tenets of the Declaration of Helsinki and was approved by Institutional Review Board/Ethics Committee of the participated hospitals.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The size of the CHAI cohort was indicated in the tables.
Data exclusions	No data were excluded after passing the initial quality-control step.
Replication	Replication was not relevant. We used independent validation cohorts.
Randomization	Samples were randomly allocated to the training, tuning and testing sets.
Blinding	All data were de-identified to remove any patient-related information.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a | Involved in the study

☒ ☐ Antibodies

☒ ☐ Eukaryotic cell lines

☒ ☐ Palaeontology and archaeology

☒ ☐ Animals and other organisms

☒ ☐ Clinical data

☒ ☐ Dual use research of concern

☒ ☐ Plants

Methods

n/a | Involved in the study

☒ ☐ ChIP-seq

☒ ☐ Flow cytometry

☒ ☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.