nature aging

Analysis

https://doi.org/10.1038/s43587-023-00536-5

Longitudinal machine learning uncouples healthy aging factors from chronic disease risks

Received: 9 May 2023

Accepted: 2 November 2023

Published online: 07 December 2023

Check for updates

Netta Mendelson Cohen[1,2], Aviezer Lifshitz[1,2], Rami Jaschek[1,2], Ehud Rinott[1,2], Ran Balicer[3], Liran I. Shlush[2], Gabriel I. Barbash[1,2] ✉ & Amos Tanay[1,2] ✉

To understand human longevity, inherent aging processes must be distinguished from known etiologies leading to age-related chronic diseases. Such deconvolution is difficult to achieve because it requires tracking patients throughout their entire lives. Here, we used machine learning to infer health trajectories over the entire adulthood age range using extrapolation from electronic medical records with partial longitudinal coverage. Using this approach, our model tracked the state of patients who were healthy and free from known chronic disease risk and distinguished individuals with higher or lower longevity potential using a multivariate score. We showed that the model and the markers it uses performed consistently on data from Israeli, British and US populations. For example, mildly low neutrophil counts and alkaline phosphatase levels serve as early indicators of healthy aging that are independent of risk for major chronic diseases. We characterize the heritability and genetic associations of our longevity score and demonstrate at least 1 year of extended lifespan for parents of high-scoring patients compared to matched controls. Longitudinal modeling of healthy individuals is thereby established as a tool for understanding healthy aging and longevity.

According to the geroscience hypothesis[1], universal aging processes are driving multiple age-related diseases. It suggests that if the underlying aging mechanism can be targeted systematically, it may be possible to promote healthy aging and increase the lifespan while lowering the occurrence of multiple chronic conditions simultaneously[1–3]. Indeed, the most common age-related diseases, such as type 2 diabetes mellitus (T2D), chronic kidney disease (CKD), cardiovascular disease (CVD), liver disease (LD) of different etiologies and chronic obstructive pulmonary disease (COPD) increase in their prevalence and intensity with age. However, the high prevalence of these diseases[4] and their gradual and progressive characteristics implies that the majority of the human population is simultaneously aging while developing chronic disease phenotypes. The results show a high degree of correlation between all the associated age-linked clinical manifestations and a lack of a clear temporal hierarchy between them. Modeling aging and disease in an unbiased fashion, and ultimately distinguishing cause and effect within the complex interplay involving healthy aging and age-related disease[5–7], remain major challenges to be addressed.

The organization of comprehensive healthcare data in electronic health records (EHRs)[8] over the last two decades holds great promise for representing patient health and disease trajectories in an age-linked and integrative fashion[9,10]. For example, we recently demonstrated how laboratory data can be normalized for age-specific and sex-specific effects to devise personalized normal reference values[11]. However,

[1]Department of Computer Science and Applied Math, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [3]Clalit Research Institute, Ramat Gan, Israel. ✉e-mail: gabi.barbash@weizmann.ac.il; amos.tanay@weizmann.ac.il

apart from smaller survey cohorts[12–15], or disease registries that focus on diagnosis codes[16,17], information on entire populations is generally available for time windows of no more than 20 years. This can cover partial patient trajectories, such as following early (for example, 25–45 years old), middle (for example, 45–65 years old) or late adulthood (for example, 65–85 years old), but still cannot portray the complete lifelong clinical history of an individual. To understand healthy aging and chronic disease risks, 20 years are often not enough. While multiple studies reported on modeling all-cause mortality and biological age using common clinical markers[18] without a longitudinal model, the impact of specific diseases on aging and the impact of aging itself on the predisposition to disease cannot be decoupled.

Given this motivation, we developed machine learning models inferring archetypical lifelong clinical trajectories from short-term clinical data. We applied the model to comprehensive data from the Clalit system[19,20] and performed validation using the UK Biobank (UKBB) and National Health and Nutrition Examination Survey (NHANES) data[13,21,22]. This resulted in a robustly replicated working model that characterizes healthy aging within the framework of lifelong age-related disease risk across several healthcare systems. Using the model, it is now possible to deconvolute patients' health records, at any age, and compute their current prospects for healthy aging, taking a more comprehensive view of their lifetime risk to develop each of five major chronic diseases. We used this new capability to identify clinical markers that are predictive of disease-free healthy aging and to reapproach heritability and genetic association of longevity-linked phenotypes. Together this shows that longitudinal models tracking patients as they age and transition from health toward disease can be readily adopted to replace static and dichotomic representations of healthy aging and common human diseases.

Results

Longevity potential in healthy individuals at 80 years of age

Multiple studies showed that short- to medium-term all-cause mortality is predictable from standard clinical data[23–26]. However, it is unclear to what extent such standard data encode longevity potential beyond the dichotomy of common age-related diseases and healthy aging. We reasoned that if longevity can be encoded by standard laboratory data in healthy individuals, we may be able to use such encoding for tracking the evolution of longevity potential across entire populations focusing on dynamics that occur long before the onset of age-related diseases. To test this, we used data on the medical histories of 4.57 M individuals from the Clalit Healthcare Services database[19] (CHSDB), providing a total of 29.5 M patient-years between ages 30 and 85, with a median tracking duration of 16.6 years (Extended Data Fig. 1). We first derived a machine learning model predicting 5-year survival for 80-year-old patients using 3 years of medical history. We found that the model recovered a sensitive spectrum of risk levels that went well beyond a simple classification of healthy individuals (Fig. 1a). For example, the model provided separation of patients on the extreme healthy end of the risk distribution, showing that the 0–2% top-scoring patients have significantly better prognosis than the second-best 2–4% scorers ($P = 0.00079$; Extended Data Fig. 2a).

To understand how standard clinical data encode an ultra-healthy state at age 80, we searched for laboratory tests that correlate with the longevity score throughout its entire range (Fig. 1b and Extended Data Fig. 2b). This highlighted clinical laboratory indicators, such as red blood cell distribution width (RDW), C-reactive protein or albumin, which showed remarkably continuous links with prognosis over their entire normal (and abnormal) ranges. Next, we asked if prognosis at age 80 is mostly a function of existing diagnoses of well-characterized age-related diseases or whether it represents more poorly defined health factors. We used the multilayered CHSDB data to identify all key diagnostic classes linked with reduced survival (Extended Data Fig. 2c) and showed that while their prevalence is correlated with the longevity score at age 80 (Fig. 1c, green), this correlation explains only part of the longevity score variance (Extended Data Fig. 2d). Moreover, for most of the major age-related diseases, we observed that incidence at ages 80–85 was uncorrelated with the overall longevity score at age 80 (Fig. 1c, purple). Together, our model defines the longevity potential at age 80 quantitatively by encoding a physiological state that may be partially driven by age-related chronic disease, but cannot be fully explained by them. The score is defined quantitatively, suggesting it can be used as a possible metric for health that extends and generalizes classification of patients into healthy and unhealthy, and motivating models tracking quantitative change in health potential rather than prediction of transition from health to disease or mortality.

Inferring longevity potential over all ages

To infer the dynamics of longevity potential across ages, we developed a machine learning algorithm for inferring a fully longitudinal model from partial patient histories. The algorithm defines a longevity potential score for each age, starting from the older age of 80 with existing mortality data. Then, in an iterative process, it derives short-term models that predict either mortality within 5 years or transition to lower or higher longevity potential after 5 years (Fig. 1d (left), Extended Data Fig. 2e and Methods), using data from at most 3 years of history. The algorithm weaves together the short-term 5-year models into a Markov model that allows summing all possible future trajectories of a patient through the longevity potential landscape (Fig. 1d (right) and Extended Data Fig. 2f). Importantly, as all scores are inferred from only 3 years of clinical history, the model can express long-term clinical outcomes from very short patient histories. As shown in Fig. 1e, the longevity scores in early to middle adulthood do not reflect an immediate mortality risk. Such immediate risks emerge only at ages 60 and older. In contrast, the scores separate individuals with high and low predicted probability of survival beyond 85 years even at age 30 (Fig. 1f). The certainty of the model's prediction for the high-scoring individuals increases with age, as more information can be extracted from standard laboratory tests. At age 65, for example, high-scoring females are estimated to have a 79% probability of surviving beyond the age of 85 years, while low-scoring females have only a 13% probability. Analysis of 10-year survival statistics (going 5 years beyond the data provided to the learning algorithm) provided initial evidence supporting the model's predictions (Fig. 2a). In addition to the change in the model's predictive value, the contribution of clinical markers to the longevity score also changed dramatically across ages (Fig. 2b). For example, alkaline phosphatase (ALP) is a strong marker in younger adulthood, while glucose and cholesterol markers impact the mid-adulthood age range and albumin and RDW the older ages. In summary, the new algorithm introduced in this study defines the long-term longevity potential of patients quantitatively and based on short-term clinical histories across early, middle and late adulthood.

Inferring shared and specific common disease lifelong risk

To map systematically lifelong disease predisposition and its contribution to our longitudinal longevity score, we defined five major chronic diseases in CHSDB, including T2D, CKD, CVD, COPD and (broadly defined) LD[4,27–30] (Fig. 3a). All of these chronic diseases are characterized by strong age-related incidence[31,32] (Fig. 3b) and are potentially driving, or are affected by, the dynamics of the longevity potential as characterized above. To study these effects systematically, we developed an extended disease risk Markov model with data describing the onset of a disease. In the extended model, the short-term (5-year) risk for developing the disease are modeled quantitatively over a spectrum of scores (Fig. 4a and Extended Data Fig. 3a), with mortality modeled as a competing risk. Using a strategy similar to the one used in the longevity model, we summed all possible trajectories of patients in the health space to compute their lifelong probability for developing the disease (Fig. 4b). This was done for each of the major disease

https://doi.org/10.1038/s43587-023-00536-5

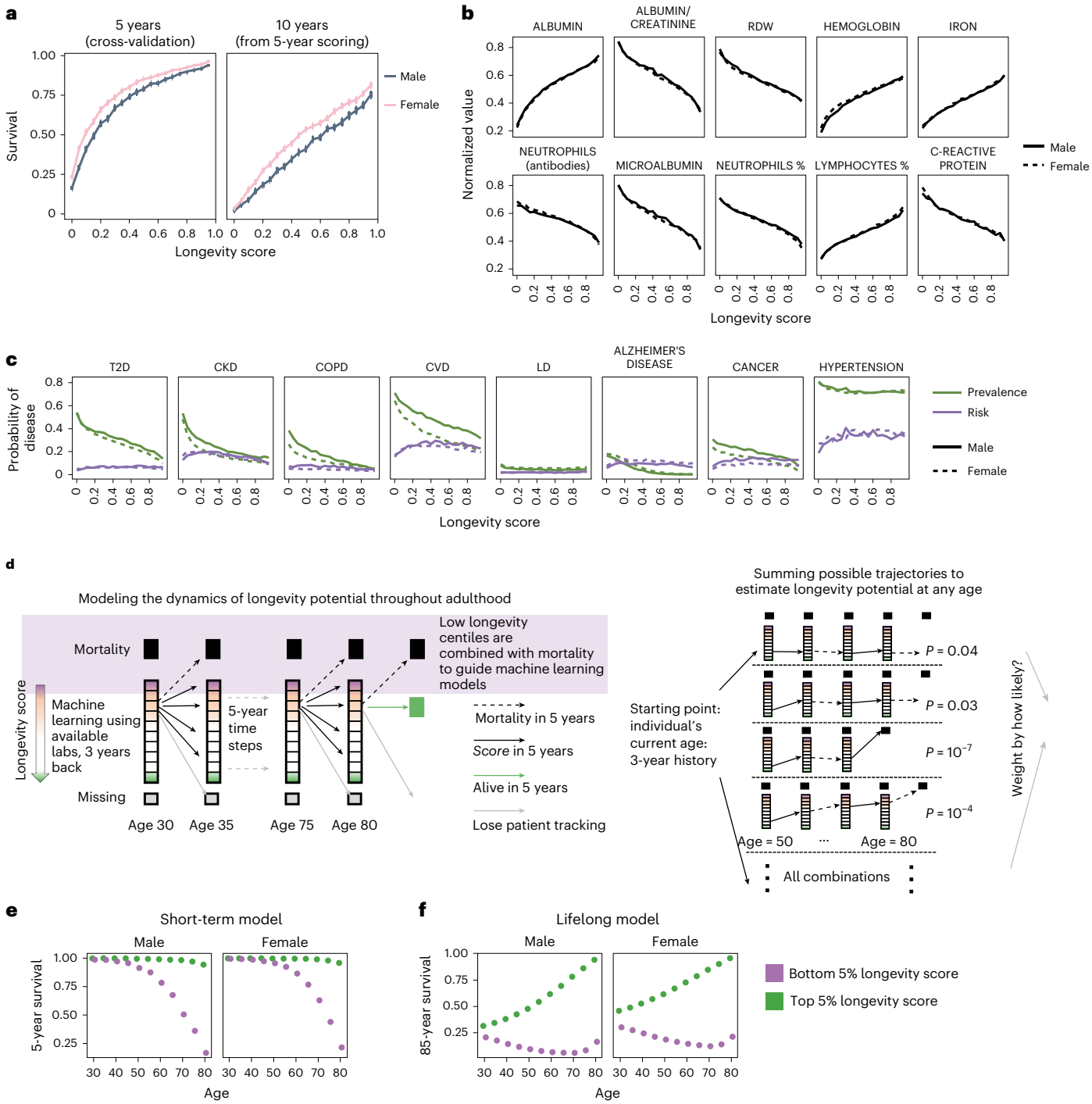


Fig. 1 | The evolution of longevity potential through adulthood. a, Sensitive longevity index for 80-year-old patients. The 5-year (left) and 10-year (right) Kaplan–Meier survival probability estimates (center points) are shown, stratified according to the computed mortality risk at age 80 (centiles, *x* axis) for all males (gray, $n = 92{,}937$) and females (pink, $n = 130{,}804$) in the CHSDB. The error bars indicate the 95% confidence intervals (CIs). Zoom-in of the top four health centiles is shown in Extended Data Fig. 2a. **b**, Standard labs encode the longevity index for 80-year-olds. The mean values of selected normalized labs according to the longevity score of 80-year-old patients are shown. **c**, Prevalence and 5-year risk for chronic disease according to the longevity score of 80-year-old patients. The fractions of patients with an existing diagnoses at 80 years of age (prevalence, green) and the fraction of patients with future diagnoses over the ages of 80–85 (risk, purple) are shown, both stratified according to the longevity potential score (*x* axis). **d**, Estimating longevity potential at any age using standard clinical data. The model defines a spectrum of longevity scores from mortality to optimal health for each age using only 3 years of clinical history per patient. Machine learning was used to predict how patients move between states in 5-year steps (left). To compute the long-term longevity potential for a patient at any age, we summed all model trajectories starting from the state encoded by the individual's short-term history (right). **e**, Five-year survival probability. The inferred 5-year (short-term) survival probabilities according to age (*x* axis) for low longevity (bottom 5%, purple) and high longevity (top 5%, green) score are shown. **f**, Lifelong survival probability. Estimation of long-term longevity potential, defined quantitatively as the probability of survival through to 85 years. Top-scoring patients (green) are identified with increasing certainty with age.

https://doi.org/10.1038/s43587-023-00536-5

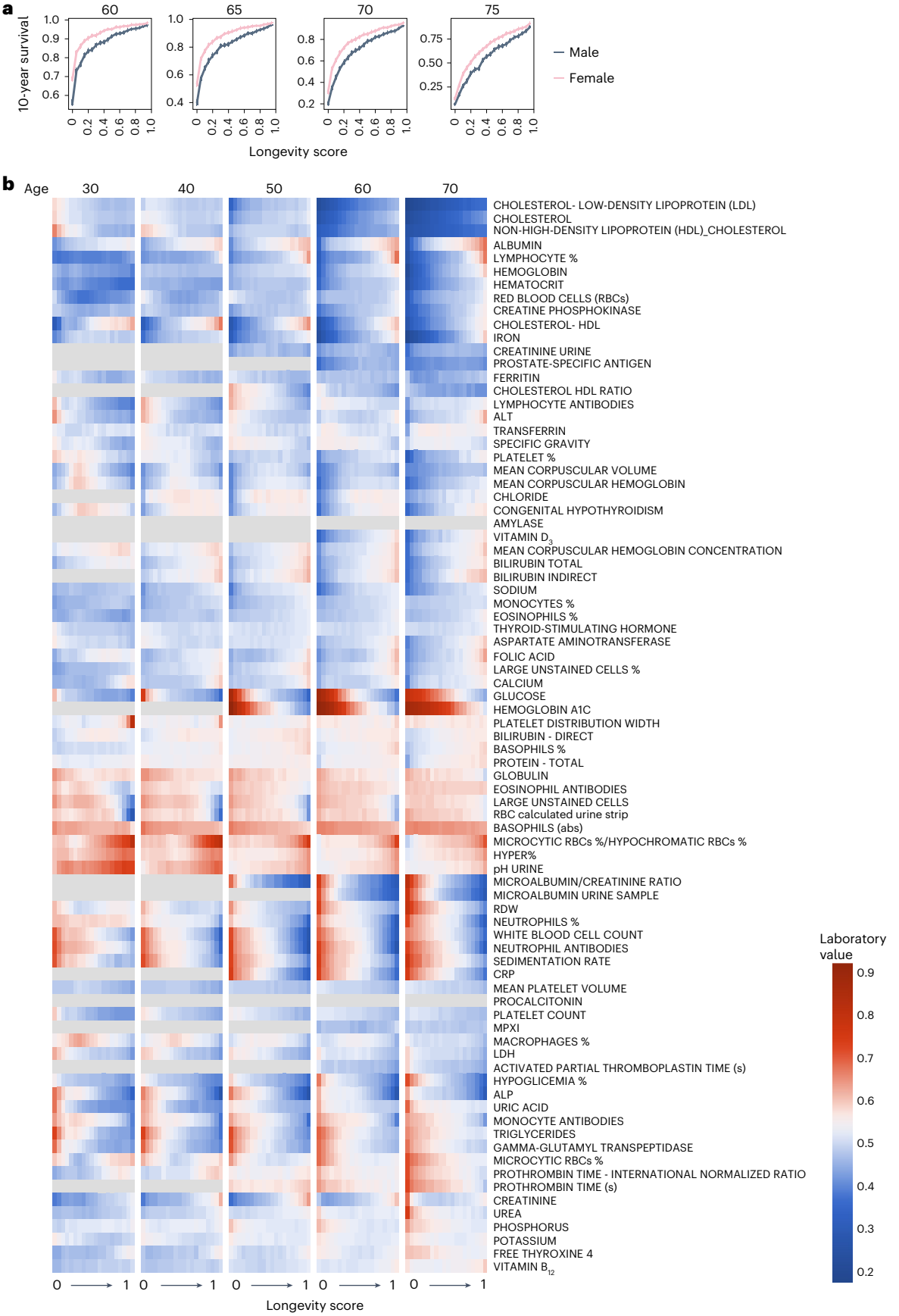


Fig. 2 | Longevity model predictive value and clinical markers contribution throughout adulthood. a, Ten-year survival according to longevity score, as in Fig. 1a for the 10-year survival according to longevity score at ages 60–75. **b**, Features of the longevity model. The mean values according to age (x-blocks) and quantile-normalized longevity score are shown. Laboratory values were quantile-normalized according to age and sex.

https://doi.org/10.1038/s43587-023-00536-5

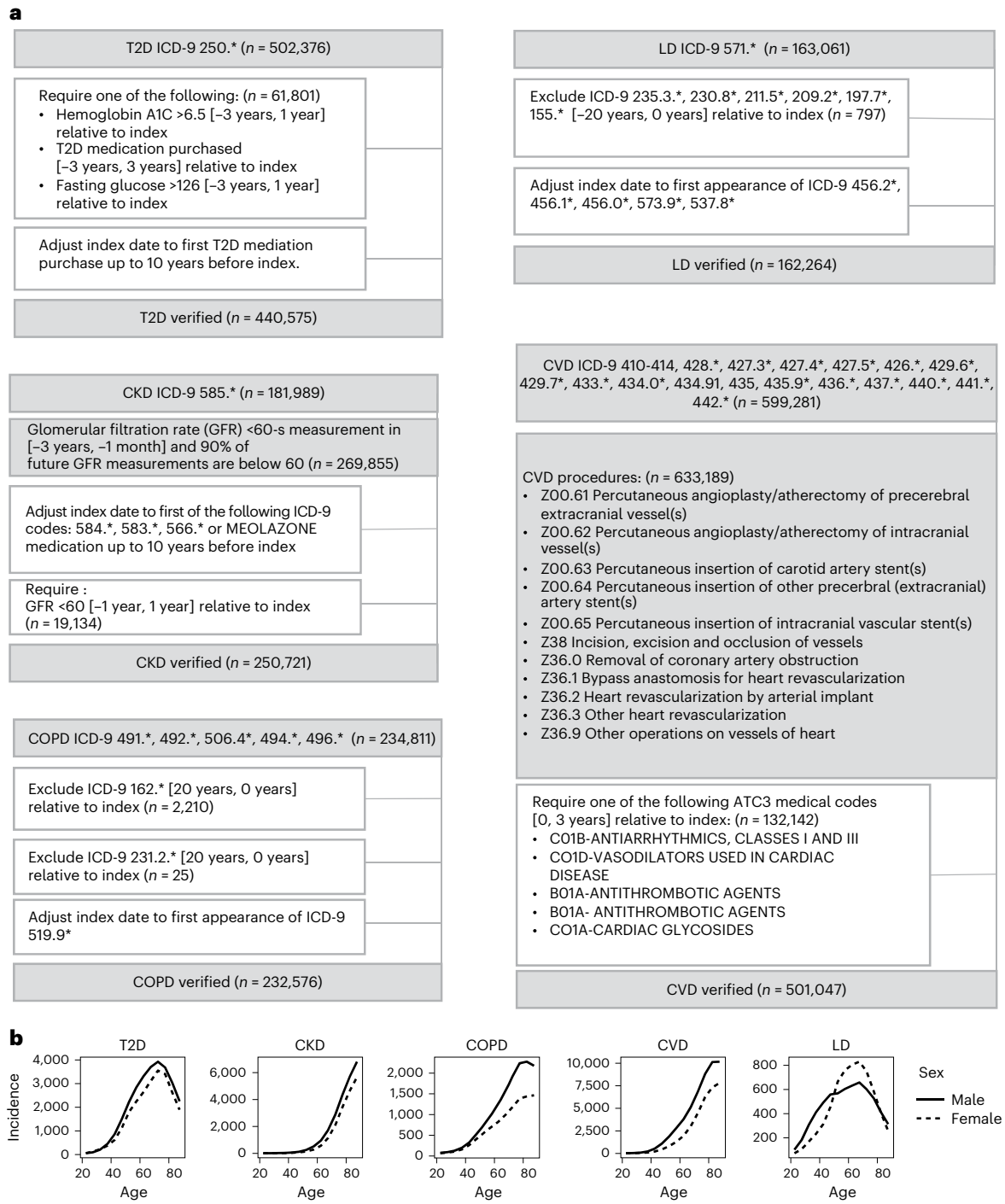


Fig. 3 | Chronic diseases. a, Cohort definitions in the CHSDB. The inclusion (gray) and exclusion (white) criteria for the disease cohorts: T2D, CKD, COPD, CVD and LD are shown. **b**, Disease incidence. The yearly incidence rates of the cohorts defined in the CHSDB are shown.

classes separately and provided us with robust separation of patients' long-term disease risks (Fig. 4c) at ages when the short-term risks are low and only weakly separable among patient groups (Extended Data Fig. 3b). The lifelong risk for a disease given a currently healthy state can decrease with age even when incidence is still high because an increasingly larger population at risk is progressing toward the disease and is no longer considered healthy. For additional major age-related disease, longitudinal modeling was less effective. Modeling patients with Alzheimer's disease was impaired by lower diagnosis reliability and lack of supporting clinical markers (Extended Data Fig. 3c). For multiple types of cancer, we observed strong short-term effects, which reduced the impact of long-term modeling (with the exception of smoking-related or alcohol-related malignancies; Extended Data Fig. 3d,e).

The predictive power of the models and the separation of high-risk and low-risk individuals (Fig. 4c, red versus blue lines) can initially increase with age, given the increased availability of relevant clinical indicators for still-healthy patients. For example, patients at high risk for T2D are identified with higher sensitivity at ages 50–60 (more than 80% lifelong risk) than at ages 30–40 (approximately 60% lifelong risk) because available glucose and hemoglobin A1C data provide more predictive power in middle-aged patients with more routine tracking and early signs of the disease. Analysis of the key features that support the

https://doi.org/10.1038/s43587-023-00536-5

estimation of lifelong disease risk and data on their age trends for high-risk and low-risk patients (Extended Data Fig. 3f–h and Extended Data Fig. 4) highlighted the common factors contributing to chronic disease predisposition. The most powerful predictive markers at younger ages were related to overweight, sugar and cholesterol, and were linked with lifelong risk (not necessarily a direct one) to all chronic diseases, albeit with different ranges. Factors linked with more specific risks were superimposed on these effects at young ages but may change in their distribution as aging progresses. For example, high creatinine and alanine transaminase (ALT) levels are linked to an elevated risk for CKD and LD, respectively, for patients at any age, as expected. More surprisingly, their correlation with the lifelong risk of other diseases is inverted in middle age (Extended Data Fig. 3f), suggesting a pleiotropic effect and justifying the detailed longitudinal modeling. As expected, estimated lifelong disease predispositions were strongly linked with one another (Extended Data Fig. 3g,h). This highlights the need in multivariate analysis to decouple the potentials of these diseases over age and characterize their contribution to the overall health state and longevity potential as quantified above.

The longevity potential cannot be fully explained by disease risk

As expected, we observed that the longevity score is strongly correlated with the lifelong probability of developing each of the five common chronic diseases we modeled (Fig. 4d). Interestingly, some of the variation in longevity scores among individuals at all ages could not be explained by a predisposition to any of these diseases. As shown in Fig. 4e, projecting patients on a health map encoding the joint distribution of the lifelong risks of chronic diseases and the longevity score clearly showed that for a subset of patients, variation in longevity potential was observed despite their low estimated risk for all modeled chronic diseases (Fig. 4f). This subset of the healthy population decreases in size with age because disease prevalence increases with age and the model can capture chronic disease risk more accurately in middle and late adulthood (Extended Data Fig. 5). Nevertheless, variation in longevity potential in these individuals is robust and represents a gap in our understanding of the healthy aging process.

Low neutrophils and low ALP correlate with longevity potential

To screen for predictive measures of longevity that cannot be explained by known chronic disease predisposition, we defined strongly healthy patients at all ages as those with a low chronic disease risk (all scores less than 0.5) and no existing diagnoses of any type of cancer. We screened for links between the longevity score and the onset of any disease in such strongly healthy individuals and identified an intriguing association with some outcomes (for example, depression; Extended Data Fig. 6a). However, the associations with documented and well-defined diseases were overall very weak (implicating a very small fraction of the patients) and could not explain the broad changes in longevity potential for the strongly healthy patient population, as represented by the longevity score. To shed more light on the possible physiological processes that underlie the longevity potential in strongly healthy individuals, we screened for correlations with clinical markers (Fig. 4g and Extended Data Fig. 6b). As expected, clinical indicators of chronic disease risk (for example, glucose, cholesterol) were uniformly normal for strongly healthy patients (as they are free from chronic disease risk) and therefore did not account for the additional variability in their longevity scores. In contrast to these indicators, we observed low neutrophil, low ALP and a high ratio between microcytic and hypochromatic red blood cells as indicative of a high longevity score in strongly healthy individuals. Interestingly, for some clinical markers, medium levels were associated with a high longevity score in strongly healthy individuals, even though their even higher levels were associated with chronic disease risk. For example, medium (but not low) body mass index (BMI), creatinine in the 60th to 50th centile (but not lower) or liver enzymes (ALT) at the 60th percentile (but not lower) were all correlated with a high longevity score, while their much higher values were linked with chronic disease risk. For some of these healthy aging indicators, in particular low neutrophil levels[33–35], earlier reports suggested a potential link with healthy aging. For other markers, some link with an effect described before as 'integrated albunemia' was likely[36]. Overall, the model we derived leads to a somewhat unexpected hierarchy of health levels in individuals who are currently considered strongly healthy. Importantly, we used patients with at least a 10-year follow-up to ascertain the models' disease and survival predictions (which were based only on 5 years of data) (Extended Data Figs. 6c and 7), showing predictive value even for individuals who were considered strongly healthy at the time of the prediction.

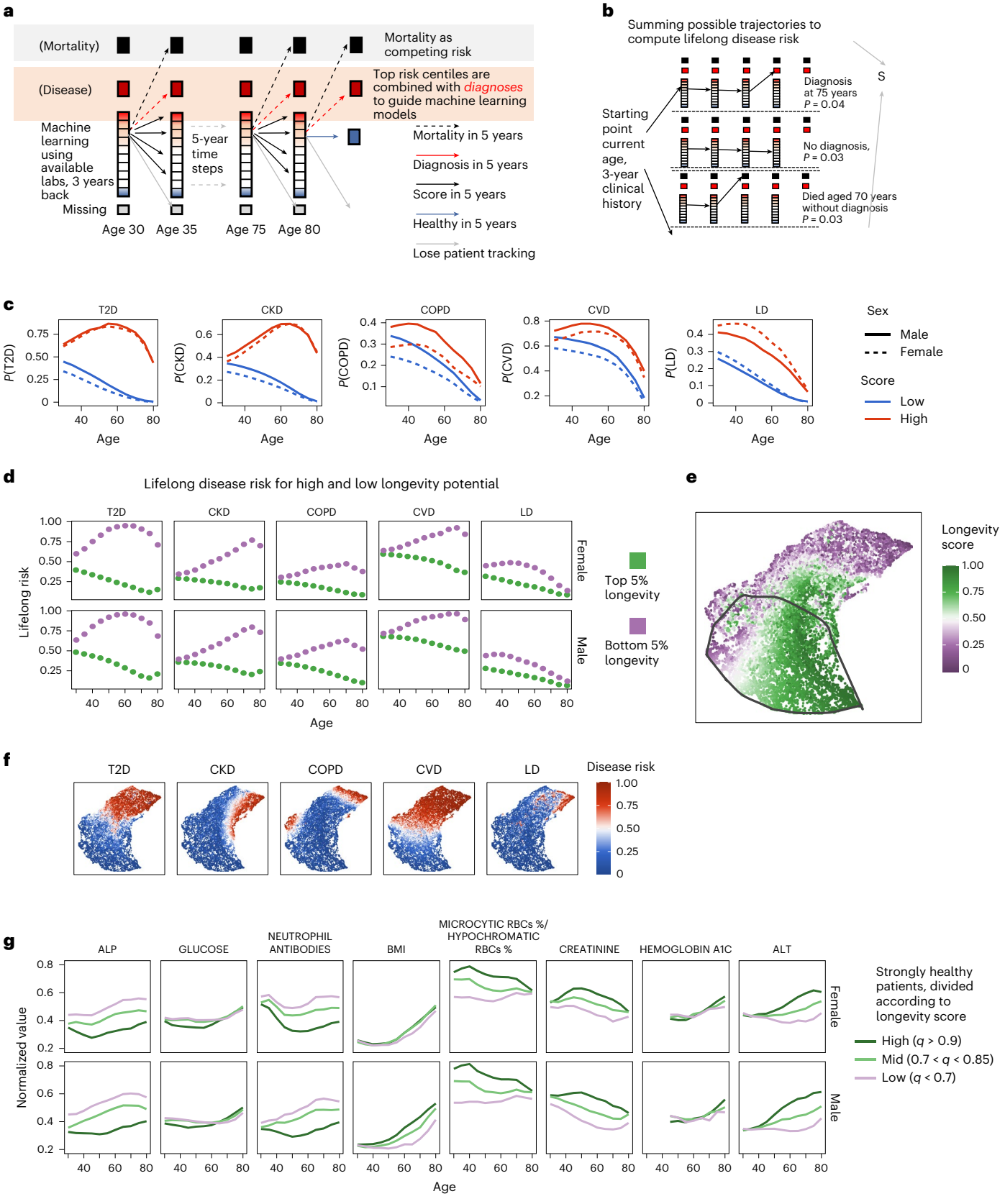
Using longevity scoring in the UKBB and NHANES populations

We used simple feature matching and renormalization (Supplementary Note 1 and Supplementary Fig. 1), but no change in model parameters, to transfer the chronic disease risks computed on the CHSDB data to the UKBB[13] and NHANES[22] population databases. The CHSDB-based longitudinal model accounts for missing data in the patients' records, allowing easier transfer to survey datasets. Also while training our longitudinal models relies on available patients' histories, applying the transferred model does not rely on longitudinal data. Our models could therefore be used to assess risks for all patients at any age between 30 and 80 years in the two cohorts. We observed highly reproducible risk stratification for chronic diseases (Fig. 5a). Analysis showed that for CVD, our scoring scheme provided risk stratification comparable (or even better, for example, for age = 65 years) to common risk metrics (Extended Data Fig. 8a,b)[37–39], even though it was not using some key features (for example, smoking history) that are not available or reliable in operational retrospective datasets such as the CHSDB one. We next transferred the longevity score to the UKBB and NHANES datasets using a similar approach. This provided both validation and an opportunity to test additional survey variables against the longevity score. We observed excellent agreement between the longevity score and 10-year survival in both the UKBB and NHANES (Fig. 5b). Key longevity laboratory correlations were reproduced in the UKBB, including low neutrophil count, low ALP and high ALT, and creatinine (Supplementary Fig. 2). As observed in the CHSDB, some of the healthy individuals in both populations varied in their longevity scores without evident

Fig. 4 | The longevity gap: modeling age-related chronic disease impact on longevity. a,**b**, Estimating lifelong disease risk at any age using standard clinical data. We extended the long-term longevity model (Fig. 1) by adding a disease of interest as the outcome and considering mortality as a competing risk. The model includes only 3 years of clinical history per patient to predict how patients move between disease risk states in 5-year steps (**a**). To compute the lifelong disease risk, we summed all possible trajectories ending with a disease state (**b**). **c**, Disease lifelong risk. Lifelong disease risk probabilities in low (bottom 5%, blue) and high (top 5%, red) disease risk according to age (*x* axis) and sex (line type). The lifelong risk for disease typically decreases at old age because the potential to become sick is proportional to the remaining number of years (to reach age 85). **d**, Lifelong disease risk stratified according to longevity. Disease model estimations are shown for patients with high and low longevity scores. **e**, The longevity space. Quantile-normalized longevity scores and disease risk for patients not diagnosed with any of the chronic age-related diseases were projected using uniform manifold approximation and projection (UMAP). Color-coded longevity scores over the projection space for patients aged 50 are shown. **f**, Disease risk on the longevity space. Like **e**, the lifelong disease risk over the longevity space has been color-coded for patients at age 50. **g**, Model features. Mean normalized laboratory values according to age (*x* axis) for key features that contribute to a high or low longevity score in strongly healthy individuals for whom the potential for all age-related chronic diseases is lower than the population median are shown.

https://doi.org/10.1038/s43587-023-00536-5

link to known diseases. The correlation of physical activity or body fat with longevity scores of healthy individuals was generally negative, while it was positive for some of the chronic diseases, as expected (Supplementary Fig. 3). A high longevity score in healthy individuals without chronic disease predisposition was almost independent of socioeconomic variables, compared to strong association of these variables with chronic disease risk (Supplementary Fig. 4). Together, the corroborated longevity scores were remarkably robust in Israeli, British and US populations, with substantial inferred predictive power for longevity in individuals lacking known disease predisposition.



https://doi.org/10.1038/s43587-023-00536-5

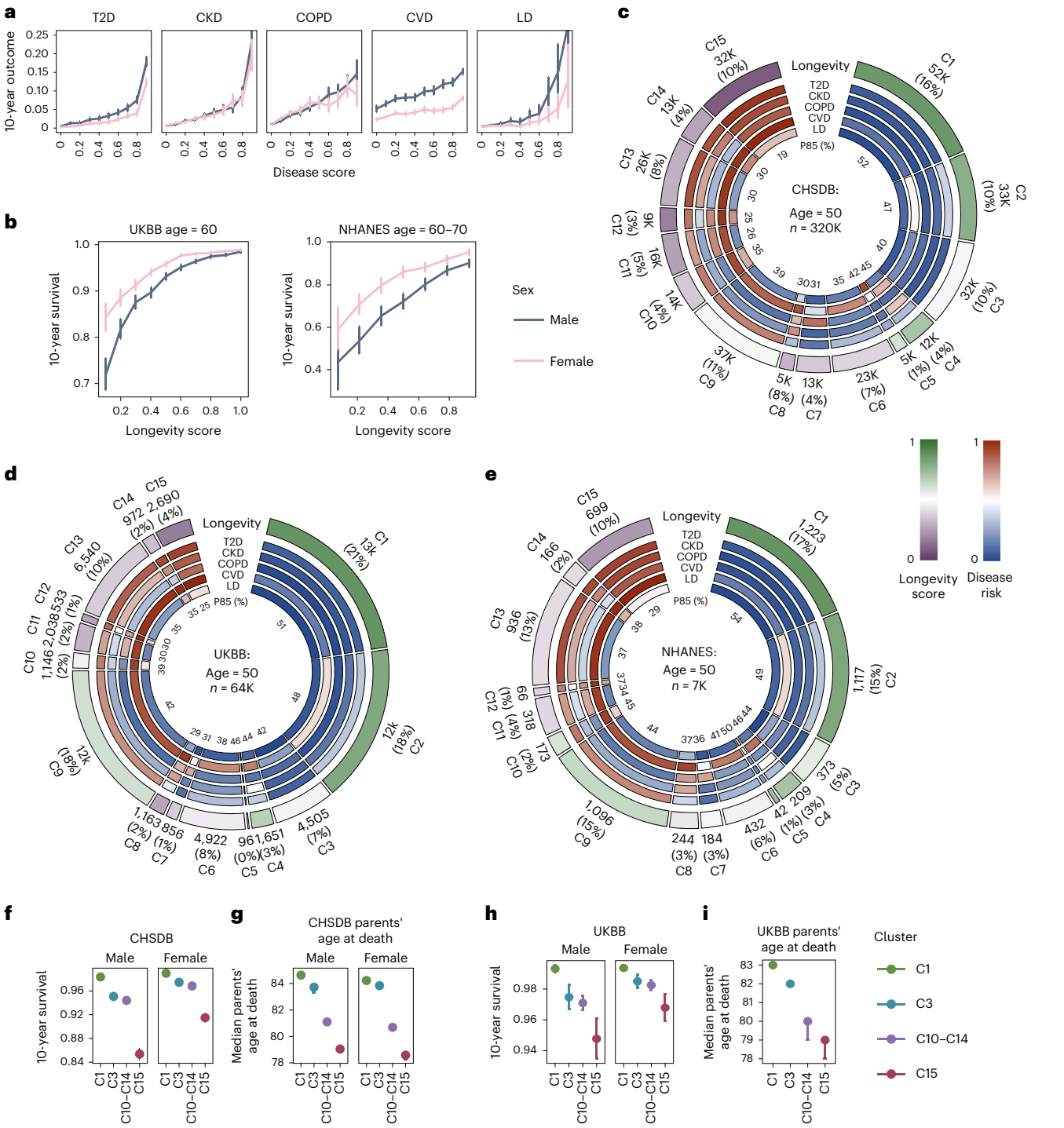


Fig. 5 | Longevity scoring in healthy individuals is robust across healthcare systems. a, Disease models performance in the UKBB. The 10-year cumulative incidence probability estimations (center points) are shown, with death as the competing risk for all patients without disease at age 60 (n = 89,124) according to the disease risk score (x axis). The error bars indicate the 95% CIs. **b**, Longevity models performance in the UKBB and NHANES. The Kaplan–Meier 10-year survival estimates for patients at age 60 or 60–70 in the UKBB/NHANES are shown according to the longevity score (x axis), n = 89,124/10,046. The error bars indicate the 95% CIs. **c**, Population distribution according to lifelong longevity and disease potential. All patients aged 50 in the CHSDB were clustered according to the quantile-normalized longevity score and disease risk. Color-coded clusters with the number of patients in each cluster (outer annotation) and probability of surviving to age 85 (P85, inner annotation) are shown. **d**, Population distribution in the UKBB (like **c** but for the UKBB data). Clustering was performed using precomputed CHSDB clusters and assigning the cluster with minimal distance to the cluster centroid. **e**, Population distribution in the NHANES (like **d** but for the NHANES data), including patients aged 50–60. **f**, Patient 10-year survival according to predisposition groups. The 10-year Kaplan–Meier probability estimates for survival (center points) for males (left, n = 81,872) and females (right, n = 110,528) are shown according to the predisposition groups shown in **c**. The error bars indicate the 95% CIs. **g**, Parental survival according to child longevity and disease potential clustering. Kaplan–Meier estimates for median age at death of the patients' parents according to the patients' predisposition groups (males, n = 27,023/14,096/54,737/23,109; females, n = 48,305/32,170/50,966/21,376) for clusters C1, C3, C10–C14 and C15) are shown. The error bars indicate the 95% CIs. **h**,**i**, As in **f**,**g** for the UKBB (n = 13,157 males and 18,765 females). Males and females combined, given a smaller sample size: n = 26,355/8,749/21,537/5,111 parents for clusters C1, C3, C10–C14 and C15.

Consistency of longevity potential across populations

To organize the multivariate disease and longevity potential across an entire population, we grouped 50-year-old CHSDB patients according to their scores (Fig. 5c and Methods), deriving 15 groups that varied from ultra-healthy (C1) to severe multi-disease risk (C15). Clustering was based only on the scores, dissecting a continuum of risks into groups for simpler analysis, but not implying that each class is completely separate from all other classes in the continuum. We then projected the UKBB and NHANES populations over the same distributions (Fig. 5d,e). The degree of strong disease predisposition at age 50 varied between populations, with 10% of the population in C15 (multivariate high risk) for the CHSDB and NHANES, and only 4% for the UKBB volunteer population (note that the latter is not an unbiased sample of the entire UK population). In all populations, we observed a robust grouping of individuals lacking any disease risk into ultra-healthy (C1, 16%, 17% and 22% of the populations) and healthy with lower longevity score (C3, 10%, 5% and 7% in the CHSDB, NHANES and UKBB, respectively). These two groups represented a universal and unappreciated variation in longevity potential of healthy individuals.

The longevity potential of patients predicts the parental lifespan

Analysis of the CHSDB and UKBB populations provided evidence for heritability consistent with the reported heritability of the clinical markers defining these risks[40–42], but low socioeconomic correlation with the longevity potential of healthy individuals (Extended Data Fig. 9a,b and Supplementary Fig. 4). As a further, stricter test for the longevity potential of healthy individuals and its heritability, we used data on parental mortality in the CHSDB and UKBB to compare the observed lifespan of the parents of individuals classified into the 15 predisposition groups outlined above. Remarkably, parents of C1 males showed a 1-year increase in lifespan in the CHSDB ($P = 2.7 \times 10^{-14}$; Fig. 5f,g) and UKBB ($P = 6.2 \times 10^{-8}$; Fig. 5h,i), compared to parents of C3 male patients. This was despite the fact that individuals in both groups lacked measurable predisposition toward the main chronic diseases. For males in the high disease risk group (C15), the decrease in the parents' lifespan was up to 5.6 years in the CHSDB and 4 years in the UKBB. Data on females' parents showed similar trends but with weaker intensity for the C1/C3 separation, suggesting that estimation of the longevity score in middle-aged females lacks precision because of perimenopausal and other effects. Evidently, because genetic transmission is incomplete and heritability only partially accounts for the longevity effects we characterized, the 1-year increase in lifespan must represent only part of the expected lifespan increase for individuals with a high longevity score. Taken together, using standard laboratory tests, the longevity score stratifies longevity potential with an impact that is associated with at least a 1-year increase in lifespan that is independent and additive to the longevity impact of known chronic diseases.

Longevity genetic association is independent of disease risk

Longevity, as the ultimate multifactorial phenotype, is linked with many genetic variants. We used our inferred predisposition groups C1–C15 in the UKBB populations to reinterpret such a link given a rich multivariate clinical context. We studied the allele frequencies (AFs) and parental survival (PS85, defined as the probability of parents to reach age 85) for 26 curated longevity variants previously defined through meta-analysis of the UKBB and multiple genome-wide association studies (GWAS)[43,44]. For each single-nucleotide polymorphism (SNP) in this set, we tagged alleles as 'good' and 'bad' according to the parental survival statistics and assigned each individual with a 'good' or 'bad' variant status accordingly (changing the homozygosity or heterozygosity criterion given the AFs; Methods). As shown in Fig. 6a, 21 of the 26 loci were indeed significantly associated with parental survival over the entire White British population (Methods). Over half (15 of 26) of the variants showed higher AFs in the chronic disease predisposition groups (C10–C15), compared to the strongly healthy groups (C1, C3), showing that these variants may be contributing to longevity, at least in part through predisposition to each known chronic disease (see Extended Data Fig. 9c for the complete C1–C15 stratification). Nevertheless, 17 of 26 variants showed a significant link with parental lifespan even when restricted to patients in the strongly healthy groups (C1 + C3). Even more interestingly, ten loci (*HLA-DQA1*, *LDLR*, *LINC02227*, *APOE*, *RAD50*, *CELSR2*, *KCNK3*, *EPHX2*, *CHRNA3*, *MICA*) showed specific enrichment in C3 patients (low disease predisposition but decreased longevity). This is suggestive of a link with longevity etiologies that are independent of the chronic disease spectrum.

Lifelong risk scores for longevity variant stratification

Beyond the support for the genetic basis of our longevity score, the data illustrate the complex interplay between the multifactorial nature of aging and the pleiotropic effects of many variants on longevity-related etiologies. In some cases, such pleiotropic effects may even be antagonistic. For example, the strongest longevity variant in the data, *APOE*, is linked with reduced parental survival within either the healthy or disease-predisposed groups. However, its AF is paradoxically lower in the disease-predisposed groups C10–C15. Longevity variants in the *APOE* locus therefore confer an overall positive healthy aging phenotype (for example, possibly because of a reported link with neurodegenerative processes) despite having a negative longevity effect by increasing the risk to common chronic diseases (Fig. 6b).

Variant association is also classically used in models aimed at inferring causal relationship in clinical features[45]; however, gaining confidence regarding such an interaction is hampered by the pleiotropic nature of most genes and the complex correlation among most clinical features. We analyzed a variant in the *ALPL* locus with strong ALP association (Extended Data Fig. 10a) to demonstrate how these challenges can be approached given our model. First, we showed that the variant is associated with reduced parental lifespan (Extended Data Fig. 10b). This may suggest a direct and causal effect through the modulation of ALP levels, but can also represent an indirect effect through any chronic disease. We computed an ALP-free longevity score from which ALP data were omitted and tested the model impact on parental survival when combined with either the genetic *ALPL* variant or the phenotypic ALP levels (Extended Data Fig. 10c). The data suggested (but provided only weak statistical support, $P < 0.018$) that the *ALPL* gene variant is correlated with differential parental survival in patients with a high ALP-free longevity score. This type of analysis may be used to infer a causal role for phenotypic ALP levels in promoting healthy aging, although the underlying mechanism is unknown.

Finally, we screened for additional variants using a GWAS on the longevity score with the five disease predisposition scores as confounders (Extended Data Fig. 10d,e). Variants identified in this way can be linked with etiologies driving longevity that are not entirely dependent (or even independent) of the chronic disease mechanisms. Using fine-mapping[46], we retained 122 loci with $P < 5 \times 10^{-8}$ (Supplementary Table 1); out of these, we selected ten variants that were not previously reported, for which the association with a lower longevity score was matched with a reduction in parental survival ($q < 0.1$; Fig. 6c). For example, we observed a variant in the *SHROOM3* locus, previously linked with neutrophil actin dysfunction and additional diseases, showing increased AF in C3 and a matching reduction in parental survival. *ADH1B*, a variant linked with addiction, showed a similar trend in patient groups C1 and C3 and a further increase in AF for the disease predisposition clusters C10–C15. Together, these analyses demonstrate the importance of taking into account a multivariate disease risk model when running and interpreting GWAS.

https://doi.org/10.1038/s43587-023-00536-5

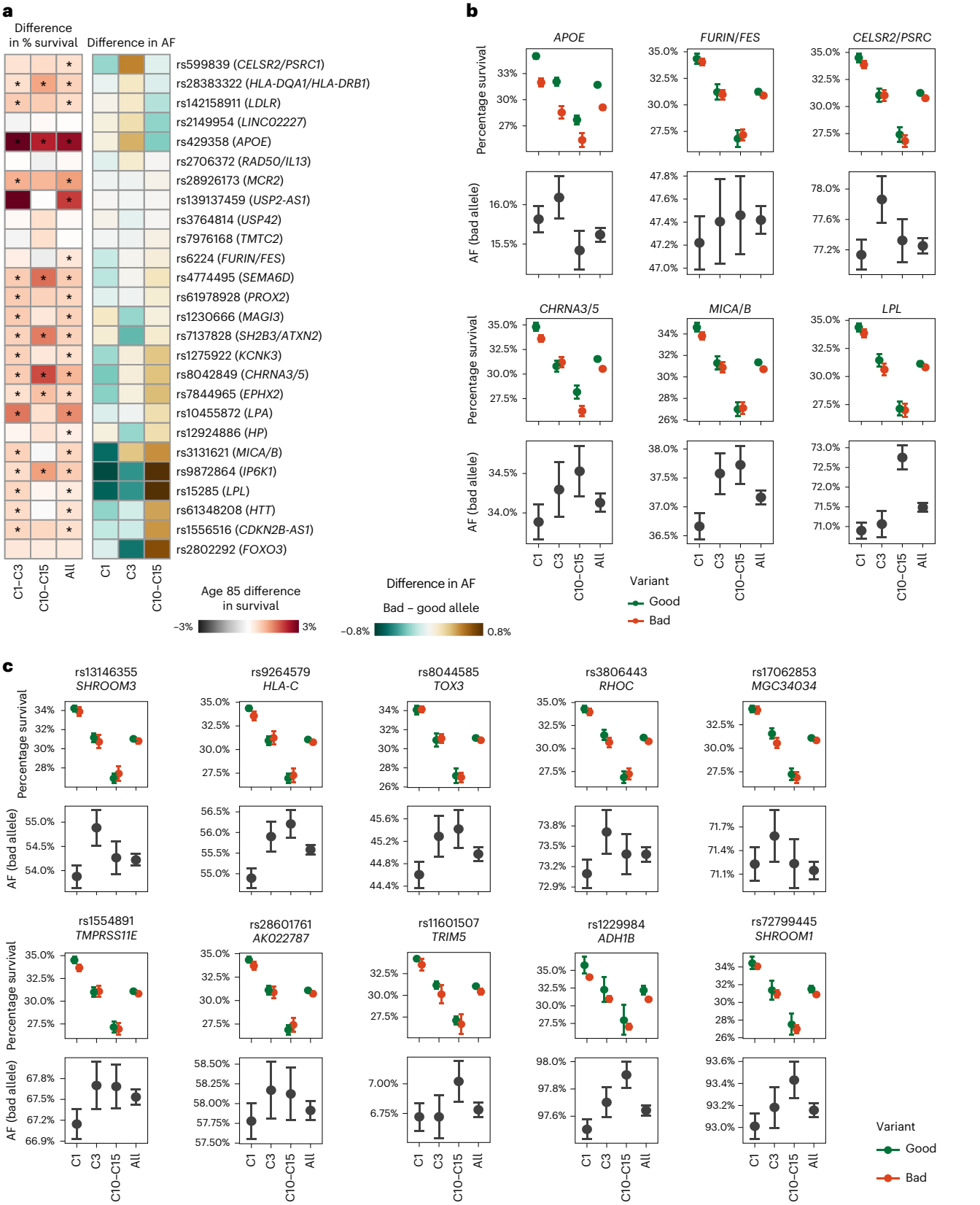


Fig. 6 | Prioritizing longevity gene variants using multivariate disease risk modeling. a, AFs and parental survival statistics (percentage survival of parents at age 85), stratified according to predisposition groups (Fig. 5), for previously described longevity variants. Data represent the difference between the 'bad' (showing poorer parental survival) and 'good' alleles. The single asterisk indicates a parental survival difference with $q < 0.05$. **b**, Parental survival (top) and AFs (for the 'bad' allele, bottom) for six selected longevity alleles. The error bars indicate the 95% CIs. **c**, We detected variants with a significant association to the patients' longevity score, with chronic disease lifelong risk scores introduced as confounder variables. Parental survival and AFs in the predisposition groups for ten of these variants are shown; they showed a consistent and significant link with parental survival. $n = 88{,}554/35{,}721/41{,}872/332{,}791$ patients in clusters C1, C3 and C10–C15, all with SNP data and parental survival (**b**,**c**). The error bars indicate the 95% CIs.

Discussion

We introduced a machine learning approach to quantify the healthy aging potential and lifelong risk for major chronic diseases in human populations. The approach is based on stitching together age-specific short-term (5-year look ahead) predictive models, each using at most 3 years of clinical history into long-term trajectory models describing aging over many decades. Focusing on healthy individuals and the dynamics of health state within them, we computed for patients at any age a longevity score expressing the likelihood of survival through to age 85. In parallel, we inferred the lifelong risk models for five key disease classes, computing for patients at any age their total probability of disease onset by age 85. By combining our models for healthy aging and major chronic diseases, we revisited classical questions on the interplay between healthy aging and chronic disease predisposition using a longitudinal and quantitative risk-based approach. We demonstrated that the models we derived based on longitudinal coverage in the CHSDB can be transferred with minimal changes to other populations in which no longitudinal coverage is available, such as the UKBB or NHANES cohorts. This makes inferences using our models readily available in essentially any modern EHR database, facilitating extension of current machine learning approaches beyond short-term predictions.

Assessing patient disease and mortality risk has been described in many studies using either classical (for example, ref. 47) or modern (for example, ref. 48) machine learning approaches. These models predict the endpoint (for example, mortality) using current patient data and suggest different solutions to account for the heterogeneity of the population and sparsity of data, and for modeling multivariate effects. The strategy in this study is reversed: rather than modeling the risk for a detrimental outcome directly, we aimed to predict the probability of maintaining the optimal health state over time. This gave rise to a model that better separates individuals with truly excellent health from normally healthy individuals and allows for the analysis of the features that correlate (and perhaps contribute) to the maintenance of such a state over prolonged time periods.

Our longitudinal multivariate model unexpectedly showed that a substantial variation in healthy aging potential can be predicted in individuals lacking any quantified risk for the spectrum of major chronic diseases. This variation was quantified using currently unappreciated signatures in common clinical markers. Interestingly, these signatures involve values well within the normal ranges[11] but provide predictive value even for young and middle-aged adults who are completely healthy based on all current definitions. This result was replicated in Israeli, British and US populations. Our model also identified all classical risk factors, such as BMI, glucose and creatinine, which are statistically linked with a potential for future development of specific diseases such as CVD, T2D and CKD. However, the additional healthy aging signatures we quantified are not (anti)correlated with any of these diseases or with future risk for their incidence while still predicting healthy aging in the long run.

After controlling for the well-described age-related chronic diseases, we aimed to understand the physiological basis of the link between healthy aging and the clinical markers we identified as predicting it. Within the complex multivariate score our model, low neutrophil levels, low ALP levels and medium (not high but also not low) levels of liver enzymes and creatinine stood out as the most predictive. A common hypothesis suggests that constitutive inflammation, which may be mediated or reported by a higher level of neutrophils, is linked with accelerated aging, providing a possible explanation for the neutrophil component in our model. The other markers are more puzzling, classifying individuals with well-balanced metabolic activity as more likely to progress well toward a healthy aging trajectory[36,49]. Poor maintenance of body weight and emergence of a frail phenotype are classical aging phenotypes for older adults. However, the predictive value of our model is already observed for relatively early ages and long before frailty is observed. This 'temporal hierarchy' suggests that these factors causally promote healthy aging rather than being caused by it.

Analysis of parental lifespan in groups of individuals classified by our model score provided a strong validation of the model's predictive power. Even when excluding all patients that are at risk for any of the age-related chronic diseases we modeled, we quantified a lifespan increase of at least 1 year for the parents of individuals with good longevity estimates compared to matched individuals with poor longevity estimates. This result is again transferred seamlessly from the CSHDB population to the UKBB cohort. Parent–child lifespan conservation was compatible with narrow sense heritability analysis of the disease risks and longevity scores in the CSHDB dataset. This supports the notion that a considerable heritable component is driving the healthy aging phenotype, beyond the well-established heritable contribution of age-related chronic disease predisposition. We suggest that genetic analysis of longevity and healthy aging can be enhanced by considering the risk models we have introduced in this study, generalizing observations in smaller cohorts manifesting an ultra-longevity phenotype[50–52] or non-longitudinal models for statistical decoupling of coarse-grained disease phenotypes[53–56]. Such genetic analysis will have to consider the remarkable pleiotropy of longevity loci, which may contribute antagonistically to disease predisposition affecting overall longevity, as demonstrated for the *APOE* locus. Considering such pleiotropy and the multifactorial nature of the aging process are essential for its understanding.

Our model's predictive performance over long timescales is still severely limited by data availability. On the one hand, there are scenarios in which more routine measurements of already accessible clinical markers, such as sugar levels in younger individuals, could improve the potency of risk estimation across the population. On the other hand, for some highly prevalent age-related diseases, the main problem is in quantifying the disease in its early stages, even before finding markers predicting its progression. This is most notably relevant for Alzheimer's disease and the dementia spectrum, on which our longitudinal model fails almost completely. This failure is most probably due to a lack of effective quantification of the disease in its early or even later stages. When the outcome is not defined, or is defined in a biased fashion, the models cannot derive effective risk estimation and extend longitudinally. As additional modalities of clinical profiling and tracking are being adopted or considered[57], for example, by adopting aging clocks[58,59] and new molecular profiling tools, we hypothesize that the potency of multivariate, multi-disease longitudinal models can increase substantially. Nevertheless, the performance of any long-range patient trajectory model, even one with unlimited access to patient profiling, is bound by the determinism of the clinical trajectory itself. In some cases, the state of a patient at a younger age is simply not predictive of later disease onset. This can be due to heavily stochastic disease etiology, as in cancer, or a strong behavioral and environmental involvement where a patient's current state is only partly indicative of their future behavior. The currently available data provide very limited socioeconomic and behavioral profiling, particularly over extended periods. Future availability of such information together with longitudinal EHRs is crucial for decoupling environmental factors, predisposition to specific chronic diseases and overall healthy aging. But even when their performance is bound, the fact that our models use all clinical data that are available in a modern healthcare system to systematically model disease risks and healthy aging prospects, provides a comprehensive estimation of the value of current clinical markers in preventive and preemptive diagnosis, and highlight the processes in which more can and should be done.

The geroscience hypothesis aims to identify aging processes that can be targeted before the emergence of age-related diseases. As shown in this study (Fig. 1), the analysis of 80-year-old individuals identified a truly remarkable spectrum of health potential established late in life and readily quantifiable using common laboratory results. The

current definitions of chronic age-related diseases account for only a small part of this spectrum, disregarding the processes that generate tremendous variation between individuals who are considered healthy by any current definition. Our computational models can capture some of this variation at younger ages. Much more work is needed to identify, characterize, monitor and treat the physiological deterioration processes that drive such variation independently of textbook diseases. Redefining a quantitative 'healthy' state is essential for enabling rational development of interventional or therapeutic approaches that will ultimately increase the probability of patients to establish and maintain a truly excellent health state throughout their adulthood.

Methods

Ethics

The CHS institutional review board approved this study (ref. no. 0158-16-COM2); the study was deemed exempt from the requirement of informed consent on the basis that the retrospective historical data were anonymized and could not be linked to any individual patient.

Overview of the Markovian lifelong disease predisposition modeling approach

The algorithm implements a reverse-forward process. It starts with the inference of a short-term model for patients at age 80 (see below the longevity prediction model for age 80), computing the risk for a clinical outcome of interest by age 85 using all available data collected from age 77 and before 80. The resulting risk scoring function can then be used for any patient in the 77–79 age range, even when the outcome in ages 80–85 is unavailable (that is, it is censored). We can then analyze patients at age 75, with clinical data collected from age 72 and before age 75, to predict a generalized outcome defined by either an observation of the clinical outcome before age 80 or the emergence of high risk for the future onset of this outcome at age 80. The latter is calculated using the model already computed in the previous stage for age 80, including all patients with available clinical data from age 77 and before age 80 (Extended Data Fig. 2e and Fig. 1d). In other words, we stretched the longitudinal scope of the data by extrapolation on inferred high-risk patients. By iterating this approach backward over the aging spectrum in steps of 5 years, we created a series of generalized outcomes representing cumulative risk, and predictive models for them for each age group. Importantly, once inferred, these models do not require more than 3 years of history for a patient at any age and can handle high levels of missing data even within this narrow time window. This concludes the reverse part of the algorithm. For the forward part, we assembled all 5-year models into a Markov chain while accounting for patients going out of coverage, missing clinical data and patient mortality. Using the composite model, we summed up the probabilities of the outcome over all age ranges, thereby estimating the lifetime outcome probability for any patient at any age, using extrapolation from a minimal clinical history (Fig. 1d right). The Markovian lifelong disease predisposition short-term prediction models and the combined Markov risk models were computed using the mldpEHR R package available at https://github.com/tanaylab/mldpEHR.

Longevity prediction model for age 80

We trained a model to classify patients according to their survival through to age 85.

Patients were included according to the following criteria: age = 80; at least one blood test in the previous 3 years (between ages 77 and 79), using white blood cell count (WBC) as a proxy; known outcome—dead within 5 years or at least 5-year follow-up. Fivefold cross-validation gradient boosting tree models were trained with features including sex, average laboratory values in the previous 3-year time window for 92 of the most common labs (Supplementary Table 3); and previous chronic diseases with a significant effect on mortality (Supplementary Table 2, previously defined in ref. 11). Patients who did not die before age 85 were considered positive cases. Training was done using the xgboost[60] v.1.2.0.1 with gbtree booster, and the binary:logistic objective using the following hyperparameters: nrounds = 1,000; subsample = 0.7; max_depth = 3; colsample_bytree = 1; eta = 0.05; eval_metric = auc; min_child_weight = 1; gamma = 0.

Inferring longevity models for age younger than 80

In an iterative process, in 5-year intervals, younger age longevity classification models were trained using the scores of the older age models (Extended Data Fig. 2e). For the model of age = x (M_x), patients were included according to the following criteria: age = x; at least one blood test in the previous 3 years (between age x−3 and x, not including), using WBC as a proxy; known outcome—dead within 5 years, or available longevity score from age model x + 5.

Patients were considered as positive cases if their score was in the top 5% of the older age model M_{x+5} and they did not die within 5 years. The training procedure was as described in the previous section (longevity for age 80).

Estimating the longevity potential

Given the age series of longevity prediction models, we constructed the Markovian lifelong longevity risk model using a reverse-forward process. The input to the risk model were the longevity scores computed for all patients in each age model, quantile-normalized according to age and sex.

State space

For each age (a) and sex (s), 20 states were defined according to the 5% quantile scores (q); an additional state was defined representing patients who were missing laboratory tests and could not be assigned a longevity score. Two possible outcomes states (for age 85) were defined: alive and dead. The Markov state space was defined as:

$$S = \left\{ S_{a,s,q} \middle| \begin{array}{c} a = \{30, 35, \ldots 80\} \\ s = \{\text{male}, \text{female}\} \\ q = \{1, \ldots 20, \text{missing}\} \end{array} \right\} \cup \{\text{alive}_{85}, \text{dead}\}$$

Starting at age 80, we computed the transition matrix $T_{80,85}$ from states $S = \{S_{80,s,q}\}$ to the two outcome states $\{\text{alive}_{85}, \text{dead}\}$ by grouping the entire population of patients aged 80 and for each state, estimating the probability for death using cumulative incidence functions from competing risk data (cmprsk v.2.2-11) that support time censoring. Continuing in a similar fashion for younger ages, we computed the transition matrices

$$T_{a,a+5=\{P(S_{a+5,s,q}) | S_{a,s,q})\} \cup \{P(\text{dead}) | S_{a,s,q}\}}$$

where we assumed that the 'dead' state is absorbing.

The probability of survival through to age 85 can then be computed if we multiply all transition matrices (from the age of interest):

$$T_{a,85} = \prod_{a}^{80} T_{a,a+5}$$

Chronic disease cohort definition in the CHSDB, UKBB and NHANES

T2D, CKD, COPD, CVD and LD were defined in the CHSDB according to the inclusion and exclusion criteria as specified in Fig. 3.

In the UKBB, chronic diseases were defined based on diagnosis codes (International Statistical Classification of Diseases and Related Health Problems, 9th (ICD-9) and 10th (ICD-10) revisions) in the showcase fields 41270 and 41271, primary care and hospital inpatient outcome data, first occurrences of medical conditions and self-reported questionnaires.

Diagnoses codes for T2D: ICD-9: 250; ICD-10: E10, E11, E12, E13, E14. Diagnoses codes for CKD: ICD-9: 585; ICD-10: N18. Diagnoses codes for COPD: ICD-9: 491, 492, 506.4, 494, 496; ICD-10: J41, J42, J43, J44, J47. Diagnoses codes for CVD: ICD-9: 410, 411, 412, 413, 414, 427.3, 427.4, 427.5, 428, 429.6, 429.7, 433, 434.0, 434.9, 435.9, 436, 437, 440, 441, 442; ICD-10: I20, I21, I23.0, I24, I25, I46.9, I48.9, I49.0, I50, I51.0, I51.2, I63, I65, I66, I67, I70, I71, I72, G45.9. Diagnoses codes for LD: ICD-9: 571, 573.3, 573.8, 573.9, 576.8; ICD-10: K70, K73, K75.9, K76.1, K76.89, K76.9, K83.5, K83.8.

Disease prediction models for age 80

A binary classification model was trained for each disease separately, to classify patients according to their disease status at age 85. Patients were included according to the following criteria: age = 80; at least one blood test in the previous 3 years (between ages 77 and 79), using WBC as a proxy; known outcome—disease within 5 years or at least a 5-year follow-up; patients not diagnosed with the disease at a younger age (younger than 80 years).

Patients were considered as positive cases if they were diagnosed with the disease before reaching age 85. The training procedure was as described in previous sections (longevity for age 80).

Inferring disease models for ages younger than 80

In an iterative process, in 5-year intervals, younger age disease classification models were trained using the scores of the older age models (Extended Data Fig. 3a). For the model for age = x, patients were included according to the following criteria: age = x; at least one blood test in the previous 3 years (between age x − 3 and x, not including), using WBC as a proxy; known outcome—disease within 5 years or available score from age model x + 5; patients not diagnosed with the disease at younger age (younger than x).

Patients were considered positive cases if they were diagnosed with the disease within 5 years or had a high score in the age model x + 5. To determine the threshold on the score to be considered, we wished to balance the number of cases observed in the immediate 5-year period with the future number of cases. To this end, we computed the expected number of patients who will be diagnosed with the disease by age 85 recursively using a time-to-event model (cmprsk v.2.2.11) with death as a competing risk to estimate the probability of becoming sick within 5 years, considering the probability of having laboratory test data:

$$N\mathrm{sick}(80, n) = n \times P(\mathrm{sick}_{85} | \mathrm{age} = 80)$$

$$N\mathrm{sick}(\mathrm{age} < 80, n) =$$
$$n \times P(\mathrm{sick\ age} + 5 | \mathrm{age}) +$$
$$N\mathrm{sick}(\mathrm{age} + 5, n \times (1 - P(\mathrm{sick\ age} + 5 | \mathrm{age}) - P(\mathrm{dead\ age} + 5 | \mathrm{age}))$$

where n is the total population size at that age. Given this estimation, we considered the N sick patients with highest risk scores as positive cases (Extended Data Fig. 3a). Model definition, features and training are as defined for the disease model age = 80.

Lifelong disease risk computation

The disease risk Markov model was computed for each disease separately, like the longevity risk computation described above. The input to the risk model are the disease scores computed for all non-sick patients in each age model, quantile-normalized according to age and sex.

State space

For each age (a) and sex (s), 20 states were defined according to the 5% quantile scores (q); an additional state was defined that represented patients who were missing laboratory tests and could not be assigned a longevity score. Four possible outcomes states (for age 85) were defined: sick; dead; sick&dead; healthy. The Markov state space was defined as:

$$S = \left\{ S_{a,s,q} \middle| \begin{array}{c} a = \{30, 35, \ldots 80\} \\ s = \{\mathrm{male}, \mathrm{female}\} \\ q = \{1, \ldots 20, \mathrm{missing}\} \end{array} \right\} \cup \{\mathrm{sick}, \mathrm{sick\&dead}, \mathrm{dead}, \mathrm{healthy}_{85}\}$$

Starting at age 80, we computed the transition matrix $T_{80,85}$ from the states $S = \{S_{80,s,q}\}$ to the four outcome states $\{\mathrm{sick}, \mathrm{sick\ \&dead}, \mathrm{dead}, \mathrm{healthy}_{85}\}$ by grouping the entire population of patients aged 80 and for each state, estimating the probability for death using the cumulative incidence functions from competing risk data (cmprsk v.2.2.11), supporting time censoring. Continuing in a similar fashion for younger ages, we computed the transition matrices:

$$T\ a, a + 5 = \{P(S_{a+5,s,q}) \mid S_{a,s,q})\} \cup \{P(\mathrm{sick}) \mid S_{a,s,q})\}$$
$$\cup \{P(\mathrm{sick\&dead}) \mid S_{a,s,q})\} \cup \{P(\mathrm{dead}) \mid S_{a,s,q})\}$$
$$\cup \{P(\mathrm{sick}) \mid \mathrm{sick})\} \cup \{P(\mathrm{sick\&dead}) \mid \mathrm{sick})\}$$

where we assumed that the 'dead' states are absorbing:

$T_{a,a+5}(\mathrm{dead}|\mathrm{dead}) = 1, T_{a,a+5}(\mathrm{sick\ \&dead}|\mathrm{sick\ \&dead}) = 1$ and that there is no transition from the 'sick' to the healthy states.

The probability of disease through to age 85 was then computed if we multiplied all transition matrices (from the age of interest):

$$T_{a,85} = \prod_{a}^{80} T_{a,a+5}$$

We can then sum the probabilities of transitioning to a sick or sick&dead state at age 85, from any state (for example, score quantile) of interest.

Handling missing data

The CHSDB provides data on patients based on their routine clinical management. Importantly, for patients older than 50 years, the frequency of sampling patients' laboratory tests is increasing rapidly. For example, for 92.9% of females and 88.1% of males, at age 60 the data provide information on at least a complete blood count in the last 3 years; for 90.6% females and 85.8% males, we also have data on liver enzymes. When training the 5-year Markovian lifelong disease predisposition score models, we considered missing data as a possible value for each parameter because it can provide information on the patient state; for example, not having a measurement may be indicative of a healthier state. Also, XGBoost supports missing values[60]. However, we omitted patients for which a WBC laboratory test was missing, assuming it to be a good proxy for the entire complete blood count panel. When combining the scores into the Markovian longitudinal model, we modeled patients without any interaction with the healthcare system as a special state (see above).

When transferring our model to any system of interest, missing features were considered in the same way as in the original CHSDB model. This may have introduced some loss of accuracy when the underlying cause for missing a sample during standard clinical management is different from the cause of missing data in a survey. However, our validations in the UKBB and NHANES suggested that this was a relatively minor effect, possibly because the level of missing data in the CHSDB is low to begin with.

Predisposition UMAP projection

For each age, we identified patients who were not diagnosed (when at the respective age) with any of the chronic diseases. We then used disease scores (T2D, CKD, COPD, CDV, LD) and quantile-normalized longevity scores, and applied two-dimensional projection using UMAP with umap v.0.2.7.0 with the following parameters: n_neigbors = 10; min_dist = 0.3 (Fig. 4).

To improve the alignment between age models, the projection was x and y axis-oriented at age older than 65 according to diabetes scores

at age 65, by minimizing the sum of squared differences between the projection and a linear reference model using x and y to predict the diabetes score.

Patient predisposition clustering

For patient predisposition clustering, we extracted disease risks (T2D, CKD, COPD, CDV, LD) and quantile-normalized longevity scores for all patients at age 50. We ran k-means clustering on these six variables into 15 clusters using tglkmeans v.0.3.4 (Fig. 5).

UKBB dataset

The UKBB recruited 500,000 volunteers in England, Wales and Scotland between 2006 and 2010 (refs. 13,61). Health data include biomarkers collected on recruitment, online questionnaires on health status and links to external health-related records, including death, cancer, hospital admissions and primary care records. Genetic data include array genotyping for practically all participants. The data used in this study were obtained from the UKBB through application no. 64658.

NHANES dataset

Data from the NHANES, a dataset of representative, noninstitutionalized US residents, included anthropometric measurements, blood biomarker levels, dual-energy X-ray absorptiometry measurements, self-reported medical history and demographics. In-depth details of the survey and sampling procedures can be found on the Centers for Disease Control and Prevention's official NHANES website (www.cdc.gov/nchs/nhanes/index.htm). Data were collected for individuals aged 25–85 from the surveys conducted between 1999 and 2016. Individuals with a missing complete blood count were excluded from the analysis. Participant death event data were extracted from the National Center for Health Statistics website (www.cdc.gov/nchs/data-linkage/mortality-public.htm), with mortality data linked from the National Death Index up to 31 December 2019.

Projection of CHSDB clustering in the UKBB and NHANES

We applied clustering as performed in the CHSDB. For cluster labeling, longevity and disease scores for each individual were evaluated with reference to the cluster centroid generated on the CHSDB population for patients aged 50. Patients were assigned to the cluster with minimal distance.

Longevity and disease scores GWAS

We used the UKBB cohort, which contains 487,203 whole-genome imputation from genotyped individuals (v.3)[13].

We restricted the analysis to 13,791,467 variants quality-controlled by the Neale laboratory (https://github.com/Nealelab/UK_Biobank_GWAS#imputed-v3-variant-qc). All quality-controlled variants had an INFO score greater than 0.8, a minor allele frequency greater than 0.001 and a Hardy–Weinberg equilibrium $P > 1 \times 10^{-10}$.

GWAS participant inclusion criteria

To define a genetically White British ancestry, we selected 338,042 participants who were both self-identified as British and were verified using a principal component analysis (PCA) of their genotypes. Briefly, we performed PCA with 20 principal components over 256,630 SNPs that were clumped using bed_clumping from the bigsnpr package (v.1.9.11)[62] and were not part of long-range linkage disequilibrium regions[63]. We then trained an xgboost[60] model to predict 'British' ancestry from the 20 principal components and filtered participants for which the model score was less than 0.5. We further restricted our analysis to participants without second-degree relatives (KING kinship $< 2^{-3.5}$) and who self-identified themselves as White British in the questionnaire (field 21,000).

Phenotype definition and GWAS

For each participant with White British ancestry, we computed the longevity score and five disease scores (CVD, CKD, COPD, T2D, LD) at the closest time point to the age of 60. For the longevity score phenotype, we removed participants older than 75 years. Each score was then inverse-rank-normalized; we used the big_unvLinReg function from the bigstatsr R package (v.1.5.6)[62] to fit a linear regression model predicting each score separately, predicting the mortality score with the inverse-rank-normalized disease scores as covariates.

Fine-mapping

We filtered the SNPs for $P \leq 5 \times 10^{-8}$ (13,177 loci) and used polyfun[46] and susieR (v.0.11.92) to perform functionally informed fine-mapping and derive the posterior probabilities for 3-Mb windows, allowing up to five causal variants per window. We then selected all loci with posterior inclusion probability (PIP) of 0.5 or greater together with the locus with the highest PIP in 1-Mb windows.

Heritability and polygenic risk scores

h_g^2 was computed using linkage disequilibrium score regression[64] on the summary statistics of the mortality score GWAS with and without disease covariates, excluding SNPs on the X chromosome.

Polygenic risk scores were computed using least absolute shrinkage and selection operator regression (bigSpLinReg function of the bigstatsr package) on every fine-mapped SNP with a PIP of 0.1 or greater.

SNP allele orientation according to parental survival

For each variant of interest, we classified homozygous and heterozygous patients and tested their parents' survival at age 85 using survminer v.0.4.9. We then classified the 'bad' allele as the one linked with a lower parental lifespan. Figure 6 compares parental survival between homozygous carriers of the 'bad' allele and heterozygous patients, or between heterozygous patients and patients who were homozygous for the 'good' allele, depending on which allele was more frequent. P values were corrected for multiple testing using the Benjamini–Hochberg method.

Statistics and reproducibility

Because this was a retrospective study, no statistical methods were used to predetermine sample sizes. Data were collected from EHRs. No data were excluded from the analyses. Models were cross-validated on the CHSDB (fivefold cross-validation) with random association of patients to each fold, controlling for uniform distribution of sex and expected outcome. External model validation on the UKBB and NHANES data included all patients.

Reporting summary

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

Data availability

UKBB data are available to approved researchers via the UKBB Research Analysis Platform (www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform). The longevity GWAS results are available at https://gwasresults.s3.ap-south-1.amazonaws.com/gwas_longevity_age_sex_covar_extended.tsv.gz. The NHANES data can be accessed at www.cdc.gov/nchs/nhanes/index.htm. Access to the CHSDB data used for this study can be made available upon reasonable request, at the discretion of the CHS, subject to an internal review by A.T. to ensure that participant privacy is protected, and subject to completion of a data sharing agreement, approval from the institutional review board of CHS and institutional guidelines, and in accordance with the current data sharing guidelines of CHS and Israeli law. Subject to receipt of the aforementioned CHS consent and subsequent approvals, data sharing will be made in a secure setting, on a per-case-specific manner, solely for the purpose of reproducing the analysis carried in the research paper, as defined by the chief information security officer of CHS. Please submit such requests to A.T.

Code availability

All model training was performed by applying the newly developed mldpEHR R package (https://github.com/tanaylab/mldpEHR) on CHDB. Laboratory normalization was conducted using the labNorm R package (http://github.com/tanaylab/labNorm). The code applied to the UKBB for models score computation, patient classification and genetic analysis is available at https://github.com/tanaylab/Mendelson_et_al_2023 and in Supplementary Software File 1.

References

1. Kennedy, B. K. et al. Geroscience: linking aging to chronic disease. *Cell* **159**, 709–713 (2014).
2. Barzilai, N., Cuervo, A. M. & Austad, S. Aging as a biological target for prevention and therapy. *JAMA* **320**, 1321–1322 (2018).
3. Kennedy, B. K. et al. Aging: a common driver of chronic diseases and a target for novel interventions. *Cell* **159**, 709–713 (2014).
4. World Health Organization. *World Health Statistics 2022: Monitoring Health for the SDGs, Sustainable Development Goals* (2022).
5. Niccoli, T. & Partridge, L. Ageing as a risk factor for disease. *Curr. Biol.* **22**, R741–R752 (2012).
6. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
7. Cohen, A. A. et al. A complex systems approach to aging biology. *Nat. Aging* **2**, 580–591 (2022).
8. Marques, I. C. P. & Ferreira, J. J. M. Digital transformation in the area of health: systematic review of 45 years of evolution. *Health Technol.* **10**, 575–586 (2020).
9. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).
10. Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat. Med.* **26**, 29–38 (2020).
11. Cohen, N. M. et al. Personalized lab test models to quantify disease potentials in healthy individuals. *Nat. Med.* **27**, 1582–1591 (2021).
12. Belding, J. N. et al. The Millennium Cohort Study: the first 20 years of research dedicated to understanding the long-term health of US Service Members and Veterans. *Ann. Epidemiol.* **67**, 61–72 (2022).
13. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
14. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
15. Deary, I. J., Gow, A. J., Pattie, A. & Starr, J. M. Cohort profile: the Lothian birth cohorts of 1921 and 1936. *Int. J. Epidemiol.* **41**, 1576–1584 (2012).
16. Siggaard, T. et al. Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nat. Commun.* **11**, 4952 (2020).
17. Viippola, E. et al. Data resource profile: nationwide registry data for high-throughput epidemiology and machine learning (FinRegistry). *Int. J. Epidemiol.* **52**, e195–e200 (2023).
18. Liu, Z. et al. A new aging measure captures morbidity and mortality risk across diverse subpopulations from NHANES IV: a cohort study. *PLoS Med.* **15**, e1002718 (2018).
19. Balicer, R. D. & Afek, A. Digital health nation: Israel's global big data innovation hub. *Lancet* **389**, 2451–2453 (2017).
20. Singer, S. R. et al. EMR-based medication adherence metric markedly enhances identification of nonadherent patients. *Am. J. Manag. Care* **18**, e372–e377 (2012).
21. Centers for Disease Control and Prevention. *About NHANES* www.cdc.gov/nchs/nhanes/about_nhanes.htm (2023).
22. Cheng, C. K.-W., Chan, J., Cembrowski, G. S. & van Assendelft, O. W. Complete blood count reference interval diagrams derived from NHANES III: stratification by age, sex, and race. *Lab. Hematol.* **10**, 42–53 (2004).
23. Avati, A. et al. Improving palliative care with deep learning. *BMC Med. Inform. Decis. Mak.* **18**, 122 (2018).
24. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
25. Beeksma, M. et al. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Med. Inform. Decis. Mak.* **19**, 36 (2019).
26. Taylor, R. A. et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad. Emerg. Med.* **23**, 269–278 (2016).
27. Almagro, P. et al. Mortality after hospitalization for COPD. *Chest* **121**, 1441–1448 (2002).
28. Asrani, S. K., Devarbhavi, H., Eaton, J. & Kamath, P. S. Burden of liver diseases in the world. *J. Hepatol.* **70**, 151–171 (2019).
29. Roglic, G. et al. The burden of mortality attributable to diabetes: realistic estimates for the year 2000. *Diabetes Care* **28**, 2130–2135 (2005).
30. Wen, C. P. et al. All-cause mortality attributable to chronic kidney disease: a prospective cohort study based on 462293 adults in Taiwan. *Lancet* **371**, 2173–2182 (2008).
31. Boehme, M. W. J. et al. Prevalence, incidence and concomitant co-morbidities of type 2 diabetes mellitus in South Western Germany—a retrospective cohort and case control study in claims data of a large statutory health insurance. *BMC Public Health* **15**, 855 (2015).
32. Pelletier, C. et al. Diabetes in Canada: facts and figures from a public health perspective. *Chronic Dis. Inj. Can.* **33**, 53–54 (2012).
33. Franceschi, C. et al. Inflamm-aging: an evolutionary perspective on immunosenescence. *Ann. N. Y. Acad. Sci.* **908**, 244–254 (2000).
34. Furman, D. et al. Chronic inflammation in the etiology of disease across the life span. *Nat. Med.* **25**, 1822–1832 (2019).
35. Zahorec, R. Ratio of neutrophil to lymphocyte counts—rapid and simple parameter of systemic inflammation and stress in critically ill. *Bratisl. Lek. Listy.* **102**, 5–14 (2001).
36. Wey, T. W. et al. An emergent integrated aging process conserved across primates. *J. Gerontol. A Biol. Sci. Med. Sci.* **74**, 1689–1698 (2019).
37. D'Agostino, R. B. et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* **117**, 743–753 (2008).
38. Hageman, S. et al. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur. Heart J.* **42**, 2439–2454 (2021).
39. Conroy, R. M. et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur. Heart J.* **24**, 987–1003 (2003).
40. Fox, C. S. et al. Genomewide linkage analysis to serum creatinine, GFR, and creatinine clearance in a community-based population: the Framingham Heart Study. *J. Am. Soc. Nephrol.* **15**, 2457–2461 (2004).
41. Ge, T., Chen, C.-Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **13**, e1006711 (2017).
42. Poulsen, P., Ohm Kyvik, K., Vaag, A. & Beck-Nielsen, H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance—a population-based twin study. *Diabetologia* **42**, 139–145 (1999).
43. Singh, P. P., Demmitt, B. A., Nath, R. D. & Brunet, A. The genetics of aging: a vertebrate perspective. *Cell* **177**, 200–220 (2019).

44. Timmers, P. R. et al. Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife* **8**, e39856 (2019).
45. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
46. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
47. Clegg, A. et al. Development and validation of an electronic frailty index using routine primary care electronic health record data. *Age Ageing* **45**, 353–360 (2016).
48. Farrell, S., Mitnitski, A., Rockwood, K. & Rutenberg, A. D. Interpretable machine learning for high-dimensional trajectories of aging health. *PLoS Comput. Biol.* **18**, e1009746 (2022).
49. Li, Q. et al. Homeostatic dysregulation proceeds in parallel in multiple physiological systems. *Aging Cell* **14**, 1103–1112 (2015).
50. Barzilai, N. et al. Unique lipoprotein phenotype and genotype associated with exceptional longevity. *JAMA* **290**, 2030–2040 (2003).
51. Schächter, F. et al. Genetic associations with human longevity at the *APOE* and *ACE* loci. *Nat. Genet.* **6**, 29–32 (1994).
52. Schoenmaker, M. et al. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *Eur. J. Hum. Genet.* **14**, 79–84 (2006).
53. Jabalameli, M. R. & Zhang, Z. D. Unravelling genetic components of longevity. *Nat. Aging* **2**, 5–6 (2022).
54. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
55. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across human complex traits. *Nat. Genet.* **52**, 859–864 (2020).
56. North, B. J. & Sinclair, D. A. The intersection between aging and cardiovascular disease. *Circ. Res.* **110**, 1097–1108 (2012).
57. Acosta, J. N., Falcone, G. J., Rajpurkar, P. & Topol, E. J. Multimodal biomedical AI. *Nat. Med.* **28**, 1773–1784 (2022).
58. Mamoshina, P. et al. Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* **9**, 242 (2018).
59. Pyrkov, T. V. et al. Longitudinal analysis of blood markers reveals progressive loss of resilience and predicts human lifespan limit. *Nat. Commun.* **12**, 2765 (2021).
60. Chen, T. & Guestrin, C. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
61. Allen, N. et al. UK Biobank: current status and what it means for epidemiology. *Health Policy Technol.***1**, 123–126 (2012).
62. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* **34**, 2781–2787 (2018).
63. Price, A. L. et al. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 132–135 (2008).
64. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

Acknowledgements

We thank N. Rappaport, A. Bercovich and O. Milman for critical reading of the manuscript and all members of the Tanay laboratory for discussions. Research at the Tanay group was supported in part by the Adelis Foundation, the Kahn Foundation, the Bolton Hope Foundation and the Israel Science Foundation BRG grant and Israel Precision Medicine program.

Author contributions

N.M.C., A.L., G.I.B. and A.T. conceived and designed the study. N.M.C., R.J., A.L., E.R. and A.T. developed the software and pipeline. R.B. provided access and initial context to the data. N.M.C., A.L. and E.R. analyzed the data with help from R.J., L.I.S., G.I.B. and A.T. N.M.C., A.L. and A.T. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

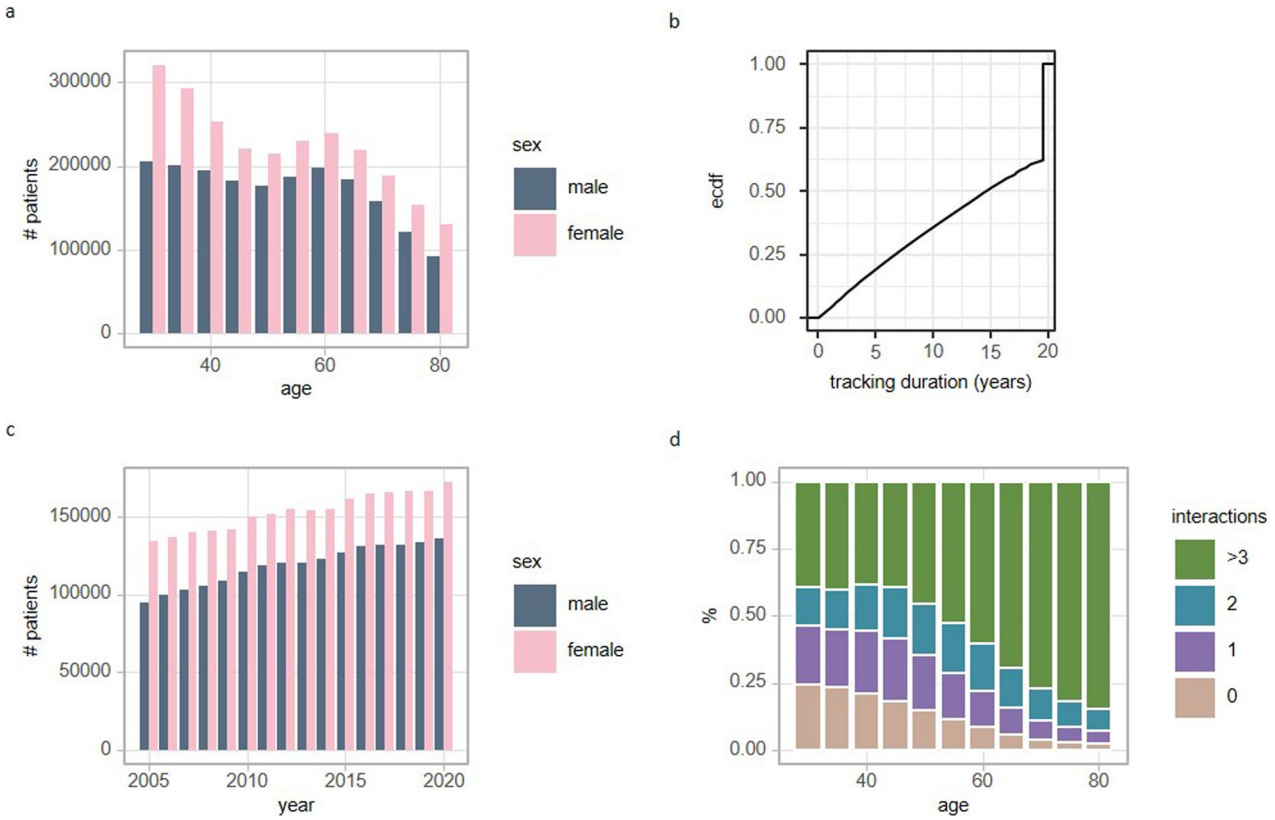
Extended data is available for this paper at https://doi.org/10.1038/s43587-023-00536-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43587-023-00536-5.

Correspondence and requests for materials should be addressed to Gabriel I. Barbash or Amos Tanay.

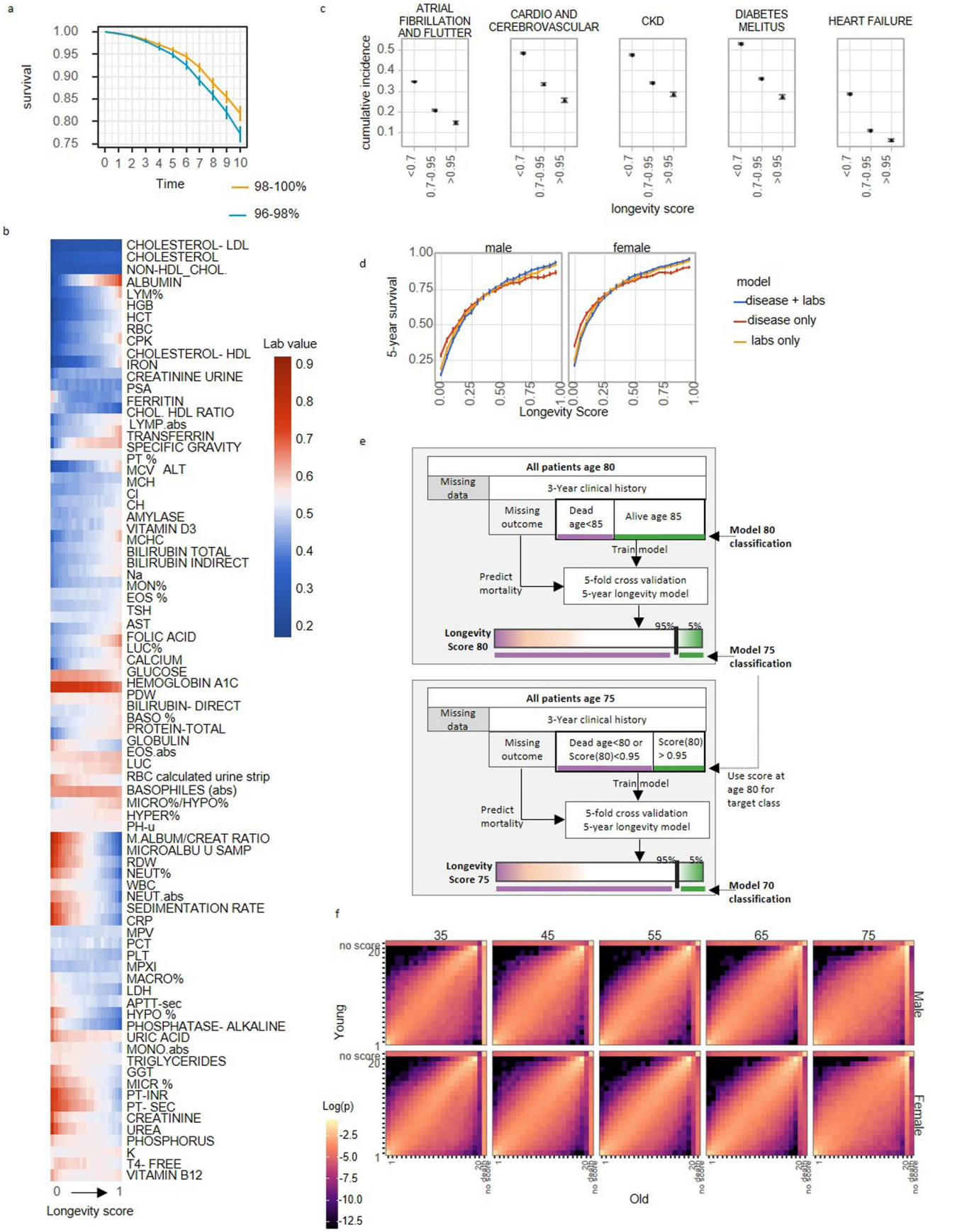
Peer review information *Nature Aging* thanks Alan Cohen, Andrew Rutenberg, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s43587-023-00536-5



Extended Data Fig. 1 | CHSDB cohort longitudinal coverage. a. Age distribution. Shown are the number of patients available for each age (x axis) and sex (grey for males, pink for females). **b. Patient tracking duration distribution**. Shown is the cumulative distribution of tracking duration (in years) for all patients in the age range of 30 to 85. **c. Chronological patient year distribution**. Shown are the number of patients, 30 < age < 85, in each calendric year. Males shown in grey, females in pink. **d. Number of lab interactions by age**. Shown is the percentage of patients at each age according to the number of WBC tests performed in the previous 3 years (as a proxy for CBC).

https://doi.org/10.1038/s43587-023-00536-5

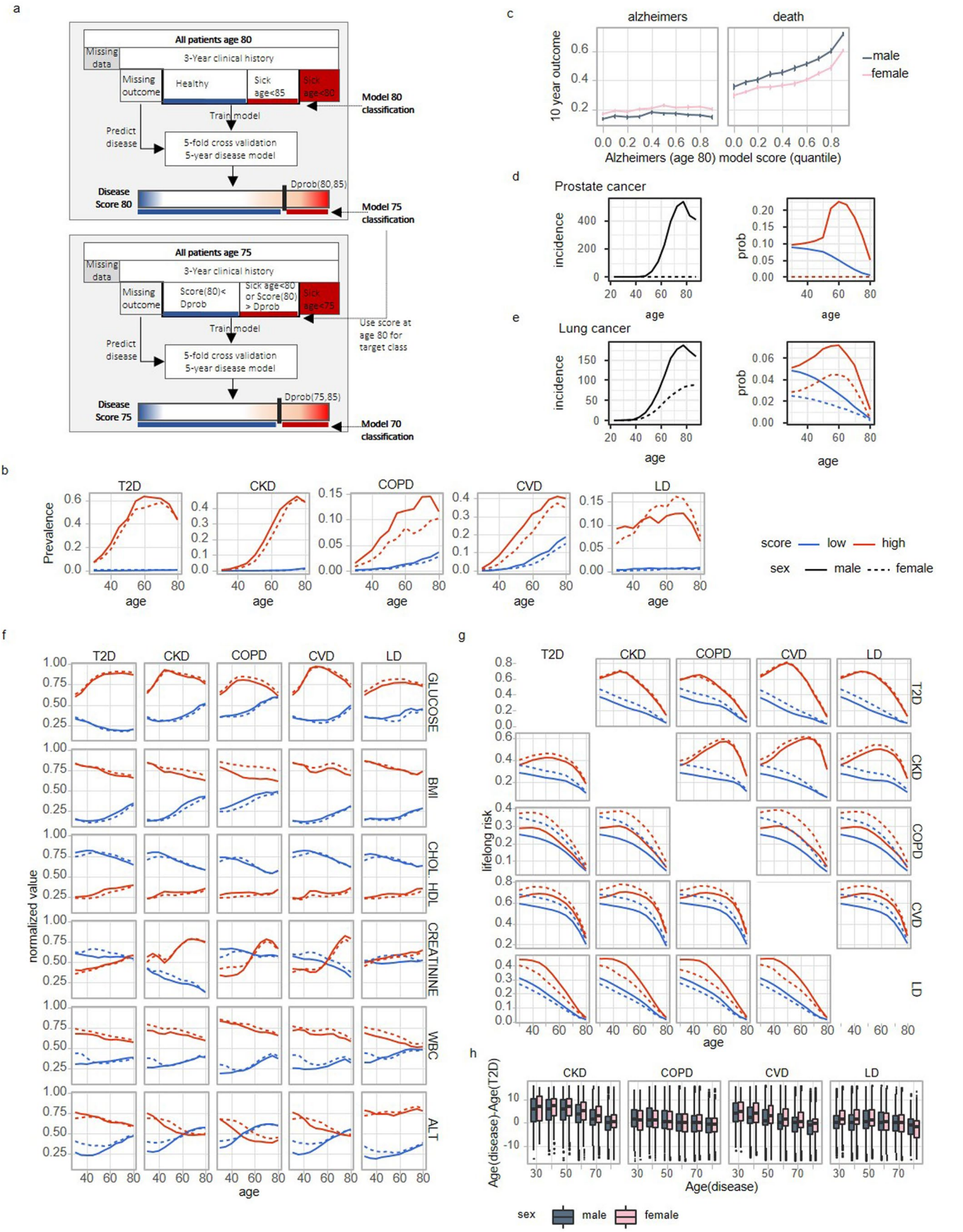


Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s43587-023-00536-5

Extended Data Fig. 2 | Longevity models. a. High longevity score survival curves. Shown are the Kaplan-Meier survival curves for the two best scoring longevity groups of patients age 80: top 98–100% (yellow, n = 4474) and 96–98% (turquoise, n = 4475). Longevity scores were computed from cross validation on CHSDB. Error bars indicate 95% confidence intervals. **b. Longevity model features**. Heatmap of mean lab (y axis) feature value for patients age 80 by quantile normalized longevity score (x axis). Lab values were quantile normalized per age/sex (11). **c. Disease predispositions by longevity**. Chronic diseases with increased mortality (11) were screened for significant difference as a function of longevity score, in estimated 8-year cumulative incidence (with death as competing risk) using cmprsk R package. Shown are the top 5 chronic diseases for patients age 80 (n = 205839), which show a significant fold-increase in disease incidence. Points (center) indicate the cumulative incidence after 8 years for the disease for each bin of longevity score (X-axis). Error bars indicate 95% confidence intervals. **d. Disease/Lab only survival model**. Shown are the 5-year survival probability estimates (center points) for males (left, n = 92937) and females (right, n = 130804) age 80 stratified by longevity score (x axis) according to a model using only disease information (red), only lab test measurements (yellow) and lab with disease data (blue). Error bars reflect 95% confidence intervals**. e. Longevity model training**. The schematic describes the longevity model training process. Starting from all patients at age 80, 5-year longevity model is trained using mortality data from patients with 3-year clinical data and 5-year known outcome, dead (0, purple) or alive (1, green). Cross validation model is applied to all patients with available clinical data, including patients with missing outcome. Top 5% scoring patients will be considered as positive cases (class=1) for age 75 longevity model. In 5-year intervals, younger age model is trained based on 5-year outcome and score from older age model. **f. Longevity model transition matrices**. Shown are the Markov model 5-year transition matrices by age (column) and sex (rows) color coded by log2 probability for transitioning from score at younger age to score at age+5 years. Score values were binned to 20 bins of 5% quantiles. An additional 'no score' bin was added for patients missing required lab data.

https://doi.org/10.1038/s43587-023-00536-5

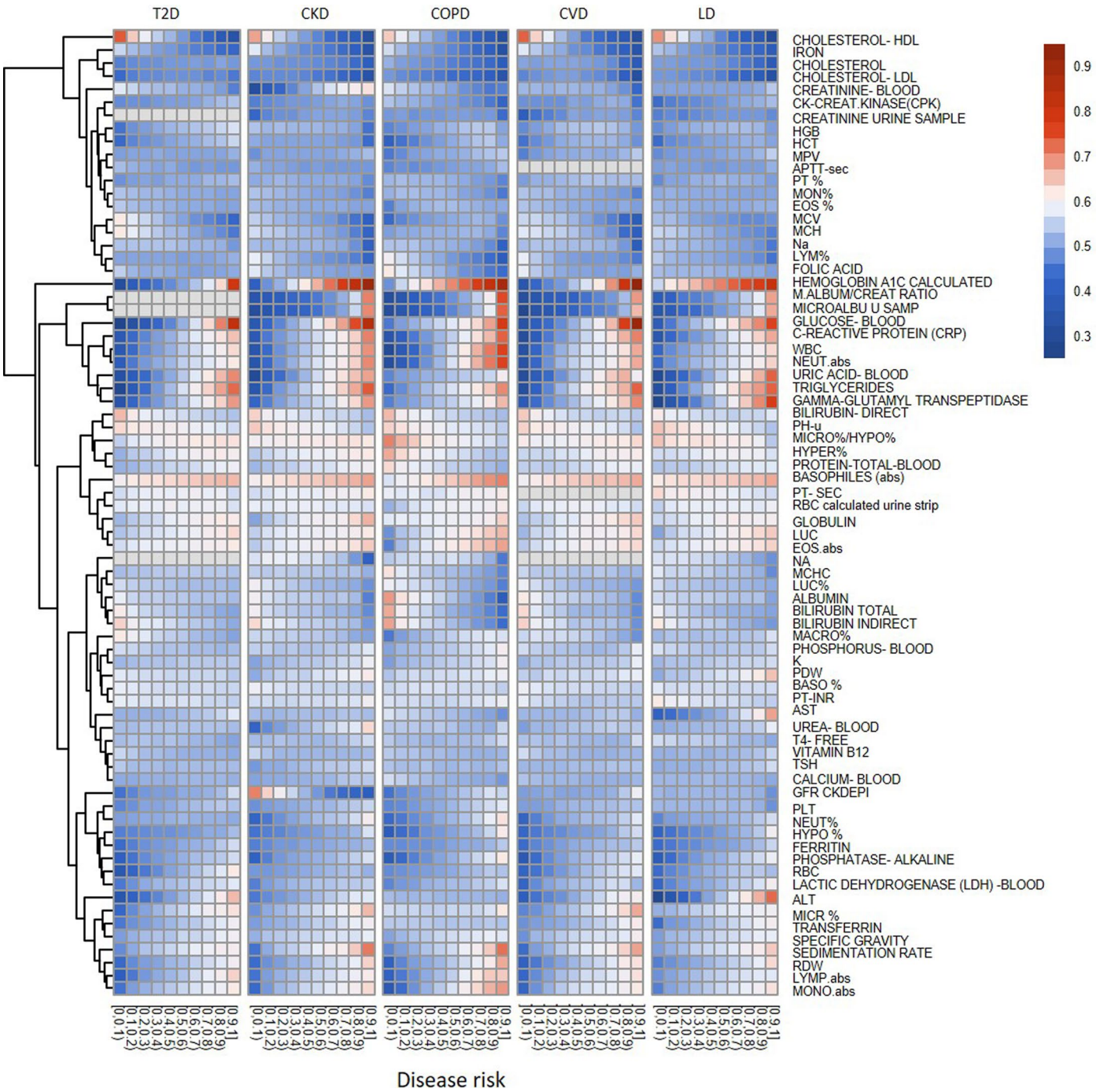


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s43587-023-00536-5

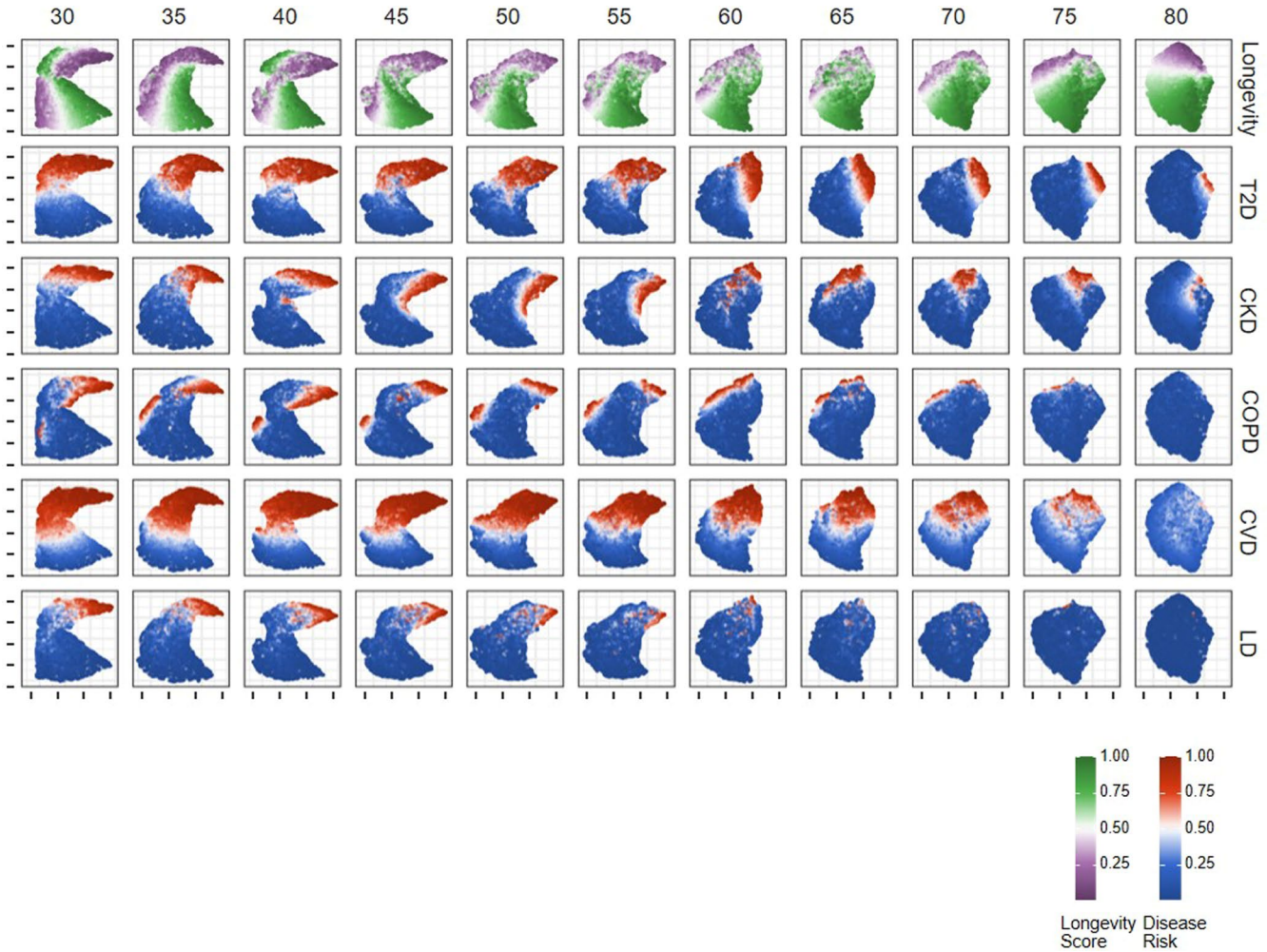
Extended Data Fig. 3 | Disease models. a. Disease model training. The schematic describes the disease models training process. Starting from all patients at age 80, 5-year disease model is trained using onset data with death as competing risk from patients with 3-year clinical data and 5-year known outcome that were not already sick at age 80, healthy (0, blue) or sick (1, red). Cross validation model is applied to all healthy (not already sick with the disease) patients with available clinical data, including patients with missing outcome. The top scoring patients, according to the expected number of patients to get sick between ages 80 and 85 computed from population cumulative incidence rates (Dprob(80,85)), will be considered as positive cases (class=1) for age 75 disease model. In 5-year intervals, younger age model is trained based on 5-year outcome and score from older age model. **b. 5-year disease prevalence**. Shown are the inferred 5-year (short-term) disease prevalence by age (x-axis) for low disease score (bottom 5%, blue) and high disease score (top 5%, red). Note that patients with suspected but unverified T2D were excluded from this analysis. **c. Alzheimer's disease**. Shown is the 10-year cumulative incidence estimation for Alzheimer's disease (left) and death (right) stratified by disease score (quantile normalized) for all patients without disease at age 80 (n = 171827) by disease score. Error bars indicate 95% confidence intervals. **d. Prostate cancer**. Shown are the incidence rate, and lifelong risk for prostate cancer disease model. **e. Lung cancer**. Similar to D for lung cancer disease model. **f. Model features**. Shown are mean values by age for key features contributing to prediction of lifelong disease risk in high- and low-risk patients (top/bottom five percentiles). **g. Relative chronic disease risks**. Similar to Fig. 4c, patients were separated into high / low risk according the disease risk listed in each column. Shown is the lifelong risk for diseases listed in each row. **h. Distribution of age difference in T2D onset and other diseases onset**. For each T2D newly diagnosed patient, we computed the time difference between T2D onset and the onset of each of the other modeled diseases (if these exist). Shown are boxplots of such time differences stratified by the age at first T2D diagnosis. The middle line indicates the median, box limits represent quartiles, and whiskers are 1.5× the interquartile range.

https://doi.org/10.1038/s43587-023-00536-5



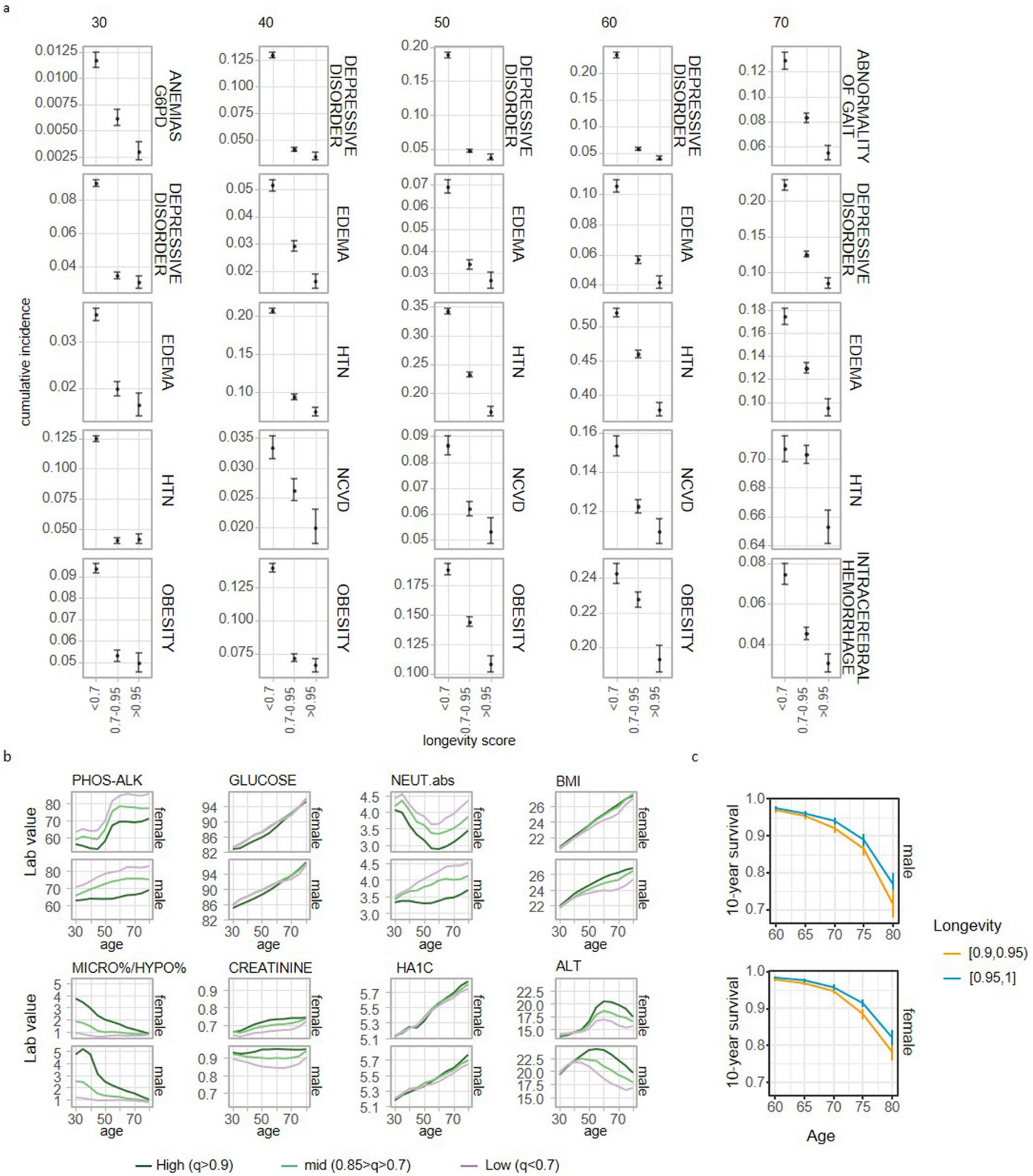
Extended Data Fig. 4 | MLDP disease models features. Heatmap of mean normalized feature value for patients age 55 for each disease stratified by quantile normalized disease risk (x-axis).

https://doi.org/10.1038/s43587-023-00536-5



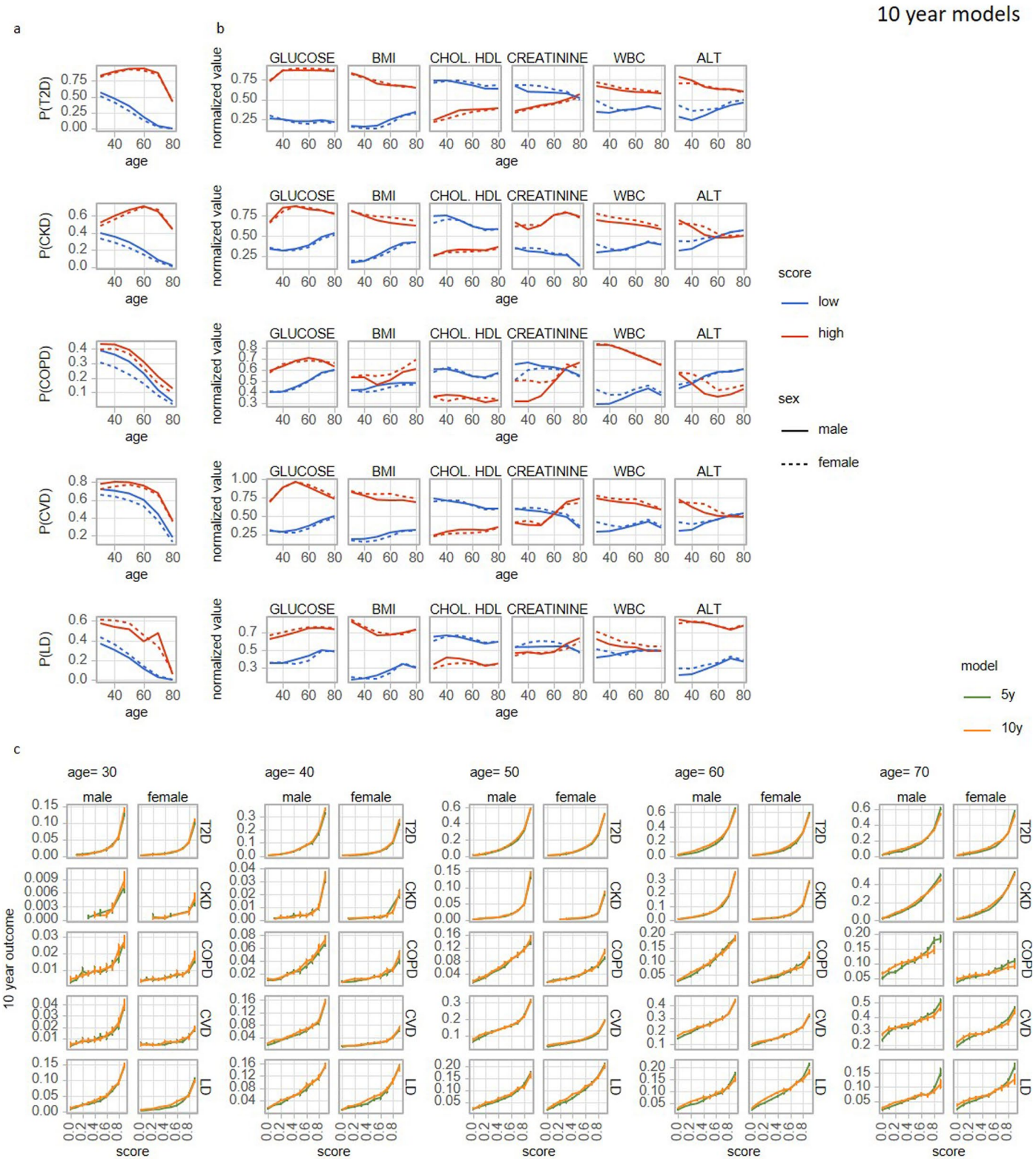
Extended Data Fig. 5 | Multivariate MLDP score UMAP projection. Quantile normalized longevity scores and disease scores were projected via UMAP for each age separately. Each dot represents a single patient. Only patients that were not diagnosed with any of the main diseases are shown.

https://doi.org/10.1038/s43587-023-00536-5



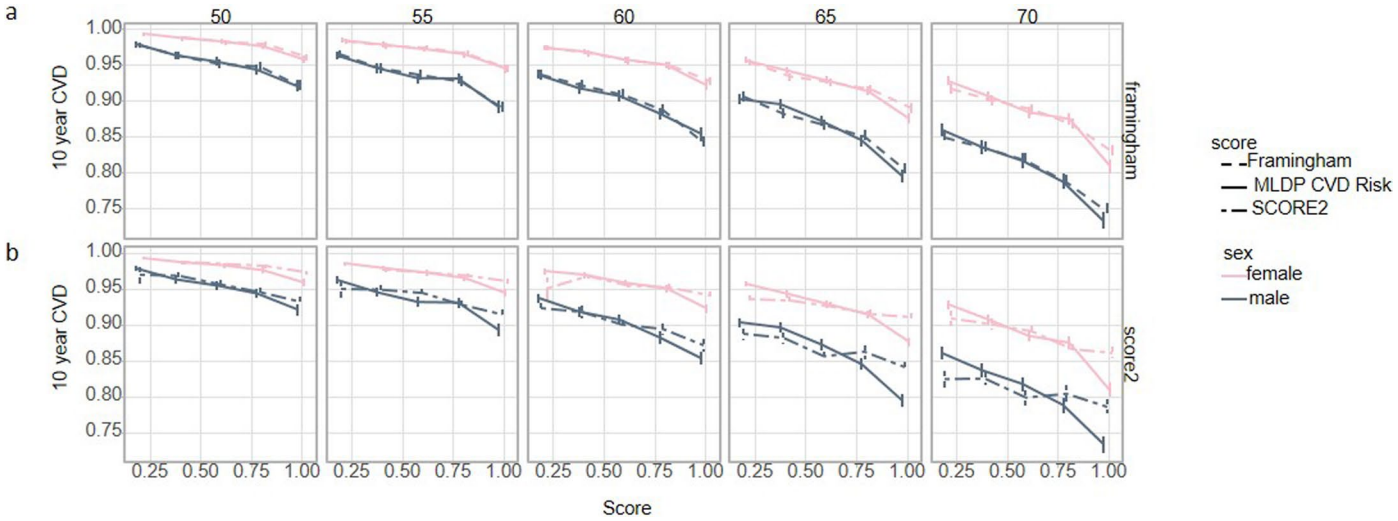
Extended Data Fig. 6 | Strongly healthy patients. a. Disease predispositions by longevity in strongly healthy patients. Similar to Extended Data Fig. 2c for patients with low disease risk score for all modeled diseases (<0.5) and that were not diagnosed with cancer. N = 157239 age 30, 134983 age 40, 91796 age 50, 86081 age 60, and 53596 age 70. Error bars indicate 95% confidence intervals. **b. Model features raw values.** Similar to Fig 4g, showing lab raw value (not normalized for age and sex). **c. High longevity score survival in strongly healthy patients.** Shown are the Kaplan-Meier 10-year survival curves for the two best scoring longevity groups: top 95–100% (yellow) and 90–95% (turquoise) longevity score by patients age (x-axis) in strongly healthy patients. N = 23168/ 21778/ 16058/ 12236/ 8872 patients at age 60/65/70/75/80. Error bars indicate 95% confidence intervals.

https://doi.org/10.1038/s43587-023-00536-5



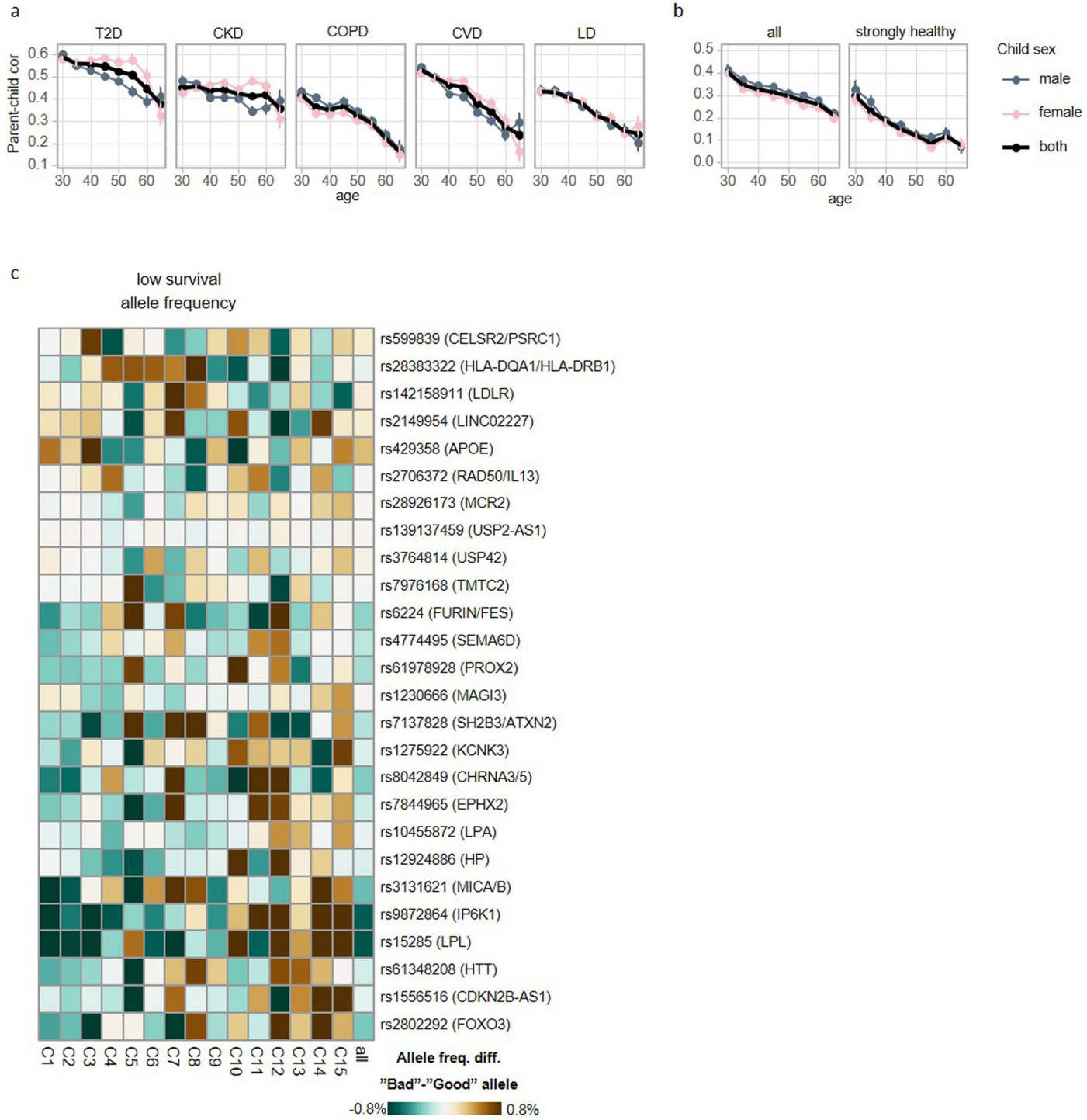
Extended Data Fig. 7 | MLDP performance using 10-year follow-ups. **a. 10-year lifelong risk models**. Similar to Fig. 4c for models trained using 10 years follow-up time. **b. 10-year Model features**. Similar to Extended Data Fig. 3f for models with 10 years follow-up time. **c. Comparing 5 y to 10 y disease model prediction**. Shown are the 10-year estimated cumulative incidence of the diseases (rows) for selected ages (columns) stratified by disease score (quantile) as computed from 5-year (n = 2176596/1935991/1597164/1789586/1469513 for age 30/40/50/60/70) and 10-year (n = 2176591/1935234/1595472/1786253/1463280 for age 30/40/50/60/70) cross validation models on CHSDB. Error bars indicate 95% confidence intervals.

https://doi.org/10.1038/s43587-023-00536-5



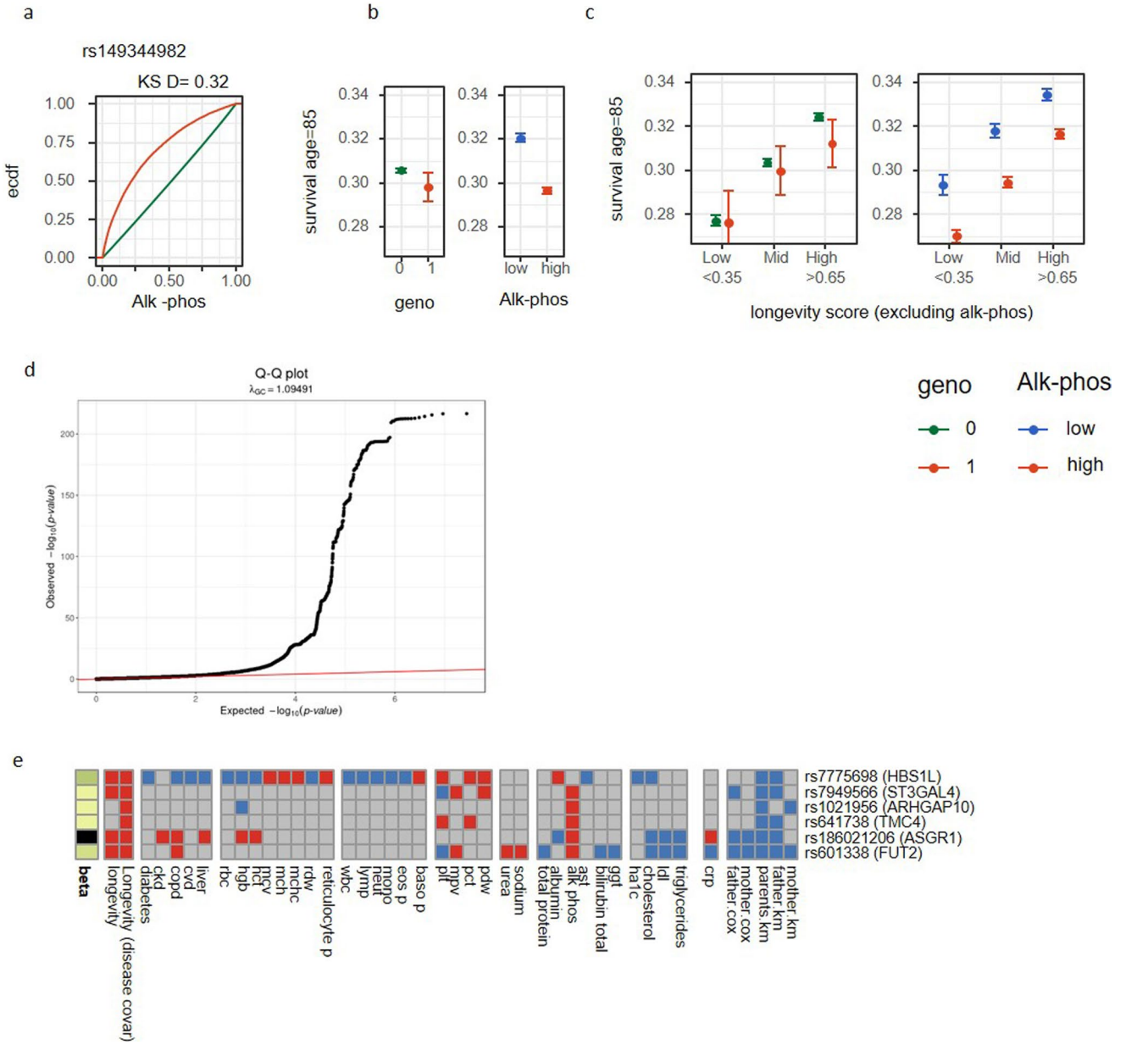
Extended Data Fig. 8 | CVD risk model validation. a-b. Framingham and Score2 score comparison. Framingham score and Score2 were computed on UKBB patients. Shown are the 10-year cumulative incidence estimates (with death as competing risk) for CVD by age stratified by score quantiles (x-axis) for Framingham and Score2 (dashed lines) and by our MLDP CVD score (solid line). N = 49486 / 55673/ 63991/ 82738/ 60391 for patients at age 50/ 55/ 60/ 65/ 70. Error bars indicate 95% confidence intervals.

https://doi.org/10.1038/s43587-023-00536-5



Extended Data Fig. 9 | Longevity heritability. a. Parent-child correlations of disease risk. Disease risks were rank-based inverse normal transformed per disease and age. Parent-child correlation was estimated as the slope of linear regression between child and average of parents whereby parents were 15 years older than offspring. Shown are the estimates of parent-child correlations for each disease ± std error, stratified by age (x-axis) computed for male offspring (grey), female offspring (pink) and combined (black). n = 430347/ 376405/ 281894/ 173863/ 93739/ 45398/ 15364/ 2321 for T2D, n = 452927/ 400504/ 303569/ 189851/ 103988/ 50925/ 17291/ 2642 for CKD, n = 451686/ 398285/ 300855/ 187115/ 101900/ 49838/ 16914/ 2575 for COPD, n = 456341/ 404196/ 306780/ 192024/ 105297/ 51745/ 17672/ 2702 for CVD, and n = 452001/ 398489/ 300724/ 186539/ 101144/ 49292/ 16734/ 2559 for LD at ages 30/ 35/ 40/ 45/ 50/ 55/ 60/ 65. **b. Parent-child correlations of longevity**. Similar to A on patients' longevity score, computed on the entire population (left, n = 536348/ 516826/ 442573/ 333617/ 238153/ 151855/ 67968/ 14587 for ages 30/ 35/ 40/ 45/ 50/ 55/ 60/ 65) and on strongly-healthy offspring individuals (right, n = 229886/ 242594/ 202371/ 141002/ 91847/ 54022/ 22286/ 4405 for ages 30/ 35/ 40/ 45/ 50/ 55/ 60/ 65). **c. Longevity snps allele frequencies**. Similar to Fig. 6a (right), showing breakdown of allele frequencies of all 15 predisposition groups as shown in Fig. 5d.

https://doi.org/10.1038/s43587-023-00536-5



Extended Data Fig. 10 | Longevity GWAS. a. Alkaline-phosphotase distribution by ALPL genotype. Shown is the cumulative distribution of alkaline-phosphotase in patients by genetic variant of ALPL rs149344982 snp. **b. Parents survival by ALPL variant and alk-phos values**. Shown are the survival estimates for parents reaching the age of 85, stratified by ALPL snp variant (left, n = 181972/4942 for genotype 0/1) and by alk-phos lab values (right, n = 70668/108379 with low/high alk-phos). Error bars indicate 95% confidence intervals. **c. Parents survival by longevity score, excluding alk-phos contribution**. Similar to B, stratifying also by longevity score, computed using a model that did not consider alkaline-phosphatase feature. N = 30945/863 for genotype 0/1 in low longevity score, n = 64518/1790 for genotype 0/1 in mid longevity score and n = 74586/1983 for genotype 0/1 in high longevity score. Error bars indicate 95% confidence intervals. **d. QQ (quantile-quantile)** plot of GWAS on the longevity score with the five disease predisposition scores introduced as confounders. Shown are observed versus expected GWAS p-values on the −log10 scale. **E. Variants with significant association to longevity score that were inconsistent with parents' survival**. Shown are the six snps, oriented by longevity score (high mortality in red), annotated by beta estimate for the longevity model with disease scores as covariates. GWAS for lab tests was obtained from Neale lab (65) and significant associations (pval < 5e-8) are depicted with red indicating variant high values and blue indicating low values inferred from GWAS beta estimates. Cox proportional hazards and Kaplan Meier survival estimates was performed on all snps on mother, father and parents combined. Also shown are Kaplan Meier (km) p-values.

nature portfolio

Corresponding author(s): Amos Tanay

Last updated by author(s): Oct 15, 2023

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 software was used for data collection.
Data analysis	Analysis was done using the R version 4.2.1 platform with the following R packages used: xgboost 1.2.01, cmprsk 2.2-11, umap 0.2.7.0, tglkmeans 0.3.4, BigSnpR 1.9.11, bigstatr 1.5.6, SusieR 0.11.92. Custom code is available in github repository https://github.com/tanaylab/Mendelson_et_al_2023

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

UKBB data is available to approved researchers via The UK Biobank Research Analysis Platform (https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform). Longevity GWAS results will be made availableare available at https://gwasresults.s3.ap-south-1.amazonaws.com/

gwas_longevity_age_sex_covar_extended.tsv.gz. National Health and Nutrition Examination Survey (NHANES) data can be accessed from https://www.cdc.gov/nchs/nhanes/index.htm.
Access to the CHSDB data used for this study may be made available upon request, at CHS's discretion, subject to an internal review by Amos Tanay to ensure that participant privacy is protected, and subject further to completion of a data sharing agreement, approval from the institutional review board of CHS and institutional guidelines and in accordance with the current data sharing guidelines of CHS and Israeli law. Subject to receipt of the aforementioned CHS consent and subsequent approvals, data sharing will be made in a secure setting, on a per-case-specific manner, solely for the purpose of reproducing the analysis carried in the research paper, as will be defined by the chief information security officer of CHS. Please submit such requests to Amos Tanay.

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	biological sex was stratified in all analysis and reported accordingly.
Reporting on race, ethnicity, or other socially relevant groupings	Race and socio-economic status were considered and reported where applicable.
Population characteristics	Medical histories of 4.57M individuals from (CHSDB), providing a total of 29.5M patient-years between ages 30-85, with a median tracking duration of 16.6 years. Demographic characteristics are described in extended data figure 1. Briefly, a total of 3.73M patients age 30-80 were included in CHSDB modeling in 5-year intervals. Mean (SD) size of each patient age group is 339518 (70120), with 44.2% (55.8%) male (female).
Recruitment	Clalit EHR, UK biobank and NHANES recruitment policies are well documented
Ethics oversight	CHS institutional review board approved this study and it was deemed exempt from the requirement for informed consent (Reference 0158-16-COM2)

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	Due to the nature of the retrospective study, no statistical methods were used to pre-determine sample sizes. Data were collected from electronic health records. No data were excluded from the analyses.
Data exclusions	None
Replication	Models were cross validated on Clalit EHR (5-fold cross validation) and then in UK Biobank and NHANES
Randomization	Cross validation group association was done while controlling for sex and outcome
Blinding	Due to the nature of this retrospective study, blinding was not applicable.

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