Causal analysis identifies small HDL particles and physical activity as key determinants of longevity of older adults

Virginia Byers Kraus,[a,1]* *Sisi Ma,*[b,j,1] *Roshan Tourani,*[b] *Gerda G. Fillenbaum,*[c] *Bruce M. Burchett,*[d] *Daniel C. Parker,*[e] *William E. Kraus,*[a] *Margery A. Connelly,*[f] *James D. Otvos,*[f] *Harvey Jay Cohen,*[d] *Melissa C. Orenduff,*[a] *Carl F. Pieper,*[d,g] *Xin Zhang,*[a] *and Constantin F. Aliferis*[b,h,i,j]

[a]Duke Molecular Physiology Institute, Duke University, Durham, NC, United States
[b]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, United States
[c]Psychiatry and Behavioral Sciences and Center for the Study of Aging and Human Development, Duke University, Durham, NC, United States
[d]Center for the Study of Aging and Human Development, Duke University, Durham, NC, United States
[e]Division of Geriatrics, Department of Medicine, Duke University School of Medicine, Durham, NC, United States
[f]Laboratory Corporation of America® Holdings (Labcorp), Morrisville, NC, United States
[g]Biostatistics and Bioinformatics, Duke University, Durham, NC, United States
[h]University of Minnesota Consortium on Aging, Minneapolis, MN, United States
[i]University of Minnesota Clinical and Translational Science Institute, Minneapolis, MN, United States
[j]University of Minnesota Department of Medicine, Minneapolis, MN, United States

eBioMedicine 2022;85: 104292
Published online 28 September 2022
https://doi.org/10.1016/j.ebiom.2022.104292

Summary

Background The hard endpoint of death is one of the most significant outcomes in both clinical practice and research settings. Our goal was to discover direct causes of longevity from medically accessible data.

Methods Using a framework that combines local causal discovery algorithms with discovery of maximally predictive and compact feature sets (the “Markov boundaries” of the response) and equivalence classes, we examined 186 variables and their relationships with survival over 27 years in 1507 participants, aged ≥71 years, of the longitudinal, community-based D-EPESE study.

Findings As few as 8-15 variables predicted longevity at 2-, 5- and 10-years with predictive performance (area under receiver operator characteristic curve) of 0·76 (95% CIs 0·69, 0·83), 0·76 (0·72, 0·81) and 0·66 (0·61, 0·71), respectively. Numbers of small high-density lipoprotein particles, younger age, and fewer pack years of cigarette smoking were the strongest determinants of longevity at 2-, 5- and 10-years, respectively. Physical function was a prominent predictor of longevity at all time horizons. Age and cognitive function contributed to predictions at 5 and 10 years. Age was not among the local 2-year prediction variables (although significant in univariable analysis), thus establishing that age is not a direct cause of 2-year longevity in the context of measured factors in our data that determine longevity.

Interpretation The discoveries in this study proceed from causal data science analyses of deep clinical and molecular phenotyping data in a community-based cohort of older adults with known lifespan.

Funding NIH/NIA R01AG054840, R01AG12765, and P30-AG028716, NIH/NIA Contract N01-AG-12102 and NCRR 1UL1TR002494-01.

Keywords: Longevity; Aging; Causal analysis; Markov boundary; High-density lipoprotein; Physical activity; Inflammation

*Corresponding author at: Box 104775, Duke Molecular Physiology Institute, 300 North Duke St, Durham, NC 27701, United States.
E-mail address: kraus004@duke.edu (V.B. Kraus).

[1] Co-first author.

Research in context

Evidence before this study

Although objective predictions of longevity are not intended to replace clinical judgement, prediction rules outperform prognostication by clinicians in some circumstances. Prior analysis of 16 validated, non-disease specific indices, led to development of ePrognosis, a test that predicts risk of mortality from 6 months to 5 years for older adults in a variety of clinical settings based on a systematic review of the literature on prognostic indices conducted through 2011. Thirteen indices used to build ePrognosis had C statistics of 0.70 or greater (validated C statistic maximum of 0.79).

Added value of this study

Using medically accessible clinical and molecular data, we identified causal determinants of longevity that achieved comparable predictivity to ePrognosis for the two- and five-year time horizon predictions requiring only 8-15 variables. The discovery of factors predicting longevity proceed from modern causal data science analyses, including predictive, local causal, Markov boundary (MB) information equivalence classes, and sepset analysis, and modelling of deep molecular phenotyping data in a community-based cohort of older adults with known lifespan. Added value is provided by the long duration of follow-up (27 years), the large number of measured variables (186); the large community-based sample (1507 individuals ≥71 years of age), and the time when the samples were obtained (1992) that provided the opportunity to evaluate 'unadulterated' lipids prior to widespread use of statins.

Implications of all the available evidence

Variables selected in every model in a Markov boundary equivalence class are guaranteed to be causal if all confounders of those variables are measured; but if unmeasured confounders of those exist, then false positive putative causes may be identified. Currently, no computational methods exist for detecting unmeasured confounders in the presence of information equivalence; consequently, this is a universal data analytic limitation. These measures clarify and enrich our understanding of mechanisms underlying longevity and could point to the most appropriate tests as indicators and focus for modification. The next stage of our (and similar) research warrants addition of dense omics assays to these data to enhance predictivity, eliminate heretofore unmeasured confounders, and reveal mediators, including druggable causal determinants of longevity.

Introduction

The complex biological networks impacting longevity are reflected in subclinical measures such as biomarkers of health and disease state. Most prior studies are limited, however, since they either examine a specialized clinical population, or only investigate a small subset of preselected variables. Moreover, cross-sectional cohort selection biases impede discovery of causal determinants of aging.[1] The **D**uke **E**stablished **P**opulations for **E**pidemiologic **S**tudies of the **E**lderly (D-EPESE) is a longitudinal cohort of community-dwelling older adults designed to overcome the above-mentioned limitations. D-EPESE included 1507 participants, aged ≥71 years with biomarker data and 27 years of death data from the time of blood sample acquisition in 1992. This research aimed to identify clinical and molecular biomarkers that predict, and causally affect, longevity, from 186 clinically accessible measures that geriatricians and clinicians can, and frequently obtain in a clinic setting.

We studied the relationships of patient-reported outcomes and questionnaires, and clinically available medical tests with survival status and identified optimal predictors. We chose to explore 2-, 5- and 10-year longevity since these time horizons are clinically relevant for this cohort of mean age 78 years with mean life expectancy of 9·37 years (men) and 10·92 years (women).[2] These time horizons are also relevant for clinical decision-making that considers the benefits and burdens of tests (e.g., colon, breast and prostate cancer screening), and treatment (stringency of lipid and blood pressure lowering), based on life expectancy.[3]

Methods

Study participants and procedures

D-EPESE is a completed longitudinal study of locally representative community-dwelling older adults established in 1986.[4] We focused on the 1,507 participants with 186 clinical and molecular measures obtained primarily at the third in-person (P3) interview in 1992 at which time blood was collected, and 27 years of death data generated since. The details of the community sampling, recruitment strategy and consent for and acquisition of blood samples are provided in Appendix S1 of the Supplementary Material.

The 186 variables consisted of sociodemographic variables including demographics/anthropometrics [age, race, sex, education, income, body mass index (BMI)], medical morbidities, self-rated health, depression, health behaviours (smoking habits, alcohol use, sleep), cognitive and physical/motor activity status, and soluble analytes including 48 traditional medical blood tests, 6 additional inflammation parameters, and 48 lipoprotein biomarkers (NMR LipoProfile® blood test). The details of the self-reported measures, sociodemographic variables, and NMR analyses are provided in Appendices S2, S3, and S4, respectively of the Supplementary Material. Summary statistics for all 186 variables for the total sample, by sex, and by time horizon are provided in Supplementary Table 1. With the exception of seven variables (interleukin-10, transforming growth factor-beta, high sensitivity vascular adhesion molecule-1, selectin, snoring, several smoking variables) with 10-30% missing, there were few missing data among the molecular and haematological measures (≤1·6%) (see Supplementary Table 1 for details). The handling of data missingness is described in the Supplementary Materials (Appendix S5).

Blood clinical chemistry and haematological measures

Venous blood was sent the day of acquisition for analyses of the SMAC-20 chemistry panel, a lipid panel (total and HDL cholesterol and triglycerides), complete blood count and serum protein electrophoresis; the remaining sample was processed and plasma transferred within 8 hours of the blood draw to long term storage at ≤-70° C. One frozen plasma aliquot was used shortly thereafter for analyses of IL-10, TGF-beta, VCAM-1, selectin, D-dimer and IL-6.[5]

Lipid and metabolite measures

The NMR LipoProfile® test was performed on non-fasting EDTA plasma aliquots that had been frozen at ≤-70°C until testing on the first thaw. Nuclear magnetic resonance (NMR) spectroscopy was performed at Labcorp (Morrisville, NC) in 2018 on a Vantera® Clinical Analyzer using the LP4 deconvolution algorithm and blinded to clinical data.[6-8] The details of the NMR analyses, performance characteristics and reference values are provided in Appendix S3 of the Supplementary Material.

Statistical and bioinformatic analysis

Predictive modelling. The outcome was dichotomous, alive vs deceased 2-, 5-, or 10-years beyond year 6 when the blood was drawn. We derived models to predict longevity at these three different time horizons using 186 features. In a discovery and cross-validation dataset (*N*=1036), we employed a nested n-fold cross validation design to obtain unbiased algorithm hyperparameters and select the best combination among several state-of-the-art classifier and feature selection algorithm families (inner loop). In the outer loop, an unbiased estimate of predictivity was obtained. In addition, we performed subsequent independent validation in a hold-out dataset (*N*=471). Performance for 2-, 5-, and 10-year longevity of the other models (age only, Levine Biological Age, and Cox models) was also computed in both the cross-validation setting and in the hold-out datasets (see Table 1). This ‘stacking” of nested cross-validation and hold-out validation, with regularization of the classifier and feature selector algorithms, provided three layers of protection against overfitting. We also compared our binary outcome models to models based on age and biological

Prediction Variable(s)	Longevity 2 years Cross-Validation	Hold-out Validation	Longevity 5 years Cross-Validation	Hold-out Validation	Longevity 10 years Cross-Validation	Hold-out Validation
ALL 186 variables	0.78±0.01	0.76 [0.69,0.83]	0.71±0.01	0.76 [0.72,0.81]	0.66±0.02	0.66 [0.61,0.71]
Markov Boundary variables*	0.72±0.01	0.74 [0.67,0.82]	0.72±0.01	0.77 [0.72,0.81]	0.64±0.01	0.67 [0.62,0.73]
Age only	0.57±0.01	0.59 [0.50,0.67]	0.61±0.02	0.62 [0.57,0.68]	0.59±0.00	0.60 [0.54,0.65]
Levine Biological Age	0.63±0.00	0.66 [0.58,0.73]	0.61±0.00	0.64 [0.59,0.70]	0.58±0.00	0.58 [0.52,0.63]

***Table 1:* Predictive performance AUCs for discriminating by longevity status at different time horizons.**
Final models indicate the best classifier and feature selection method for an outcome determined by the inner loop of the nested cross-validation. Predictive performance was estimated from cross-validation (mean and standard deviation of the estimated AUC listed) and hold-out validation (AUC and the 95% confidence interval listed). *The number of variables in the Markov boundaries were 8, 13, and 11 for 2-, 5- and 10-year longevity prediction, respectively; this subset of variables was determined by GLL. The total signal for the longevity outcome is captured by the Markov boundary indicated by the similar performance. Random Forest yielded the optimal model performance for 2- and 5-year longevity predictions with all 186 variables and the Markov boundary variables. Boosted Tree was optimal for 10-year longevity predictions with all 186 variables and the Markov boundary variables. Logistic Regression was the best model for all remaining conditions.

age, previously reported in the literature,[3,9] and time-to-death Cox proportional hazards model.

Markov boundary (MB) analysis. To identify predictive variables with likely causal interpretability for longevity, we employed MB analysis (see Appendix S6 of the Supplementary Materials for details). A MB[10] of an outcome variable is a non-reducible variable set that renders all other variables independent of the outcome, and therefore contains maximum information regarding the outcome and maximum parsimony. We used our state-of-the-art algorithms to identify the MBs.[11] According to causal graph theory, MB variables are direct causes (under the assumption of no unmeasured confounding and given that the predictive variable is a terminal one), but multiple MBs can exist in certain distributions, creating multiple causal candidate sets. To address this, we extracted all MBs for each time horizon using our validated TIE* algorithm[10] to identify all possible measured direct causal candidates and estimate ranges of their causal effect estimates using Pearl's do-calculus,[12] which guarantees unbiased and complete causal effect estimation. In addition, we applied a Sepset analysis[13] that explains why some previously reported longevity risk factors are rejected from our final models because variables in our models either mediate or confound these previous factors.

The analysis methodology employed in the present study outputs local causes of Longevity: {A, B, K} and its equivalent {A', B, K} (See Supplementary Materials and Figure 1 for more details). These sets are all Markov boundaries thus have optimal predictive signal for Longevity and are minimal (i.e., maximally compact). They can be thus readily used to create optimal predictors. Except for unmeasured confounding of variable K, the methods used successfully avoid pitfalls and challenges outlined in Figure 1; of note, no non-experimental method exists to avoid this problem. We provide a range of potential causal effect estimations for all variables in the multiple MBs. The effect estimation was computed by fitting a logistic regression model using the variables in each MB variable set as the independent variables and the corresponding outcome as the dependent variable. Pearl's do-calculus guarantees that in the absence of confounders, this causal effect estimation procedure is unbiased. For ease of interpretation, we describe variables in MBs as potentiating or attenuating longevity meaning that a higher value of the variable is associated with longer or shorter life, respectively.

Ethics

Written informed consent was provided by all participants. Subsequent annual approval was provided by Duke University Institutional Review Board (approval number Pro00010226).

Role of the funders

Funding only. Funders had no role in study design, data collection, data analyses, interpretation, or writing of the report.

Results

Characteristics of the D-EPESE sample

Clinical data for these analyses were ascertained in person (P) from baseline (P1), 3 (P2), and 6 (P3) years after enrolment, and death data from National Death Index (NDI) searches through December 31, 2019. We analysed the 1507 participants with both death data and available plasma samples from among the 1554 participants with banked samples (Figure 2). At the time of blood acquisition (P3), participants were 71-102 years of age (mean age 78, SD 5, median 77); 65% were female; 47% were white, and 53% were black (oversampled to improve statistical precision for this group as previously described[14] and detailed in Supplementary Materials). The incidence of all-cause mortality at 2, 5 and 10 years was 11·8%, 31·3%, and 71·7%, respectively. The mean time to death was 7·86 years (median 6·94 years).

Clinical and molecular variables predicting longevity

In univariable analyses, several variables were consistently associated [$P<0·05$ (t-test for continuous variables, Chi Square for categorical variables)] with longevity (survival) at all three time horizons (Supplementary Table 1 with factors associated with longevity highlighted are in bold red font). The strongest among these were Instrumental Activities of Daily Living (IADL) motor and cognitive items able to be performed, small high-density lipoprotein (HDL) particle numbers and Apolipoprotein A1 (a component of HDL). Based on standardized effect estimates, age was the strongest predictor of 5-year longevity, but pack years of smoking and ability to perform heavy housework had stronger effects for predicting 10-year longevity. Interestingly, none of the self-reported medical conditions (heart attack, stroke, cancer, diabetes, high blood pressure), even those associated with longevity in univariable analyses (heart attack, stroke, diabetes) were identified as proximal (i.e., local) causal variables of 2-, 5- or 10-year longevity, indicating that the effects of these factors are mediated or confounded by variables in the local causal models (such as small HDL particles), or their antecedent causes (such as smoking).

Causal predictors of 2-, 5- and 10-year longevity

The prediction of 2- and 5-year longevity using all 186 features was strong with area under the receiver operating characteristic curves (AUCs) of 0·78±0·01 and 0·71±0·01, respectively, as estimated by nested repeated cross-validation, and 0·76 [0·69,0·83] and 0·76 [0·72,0·

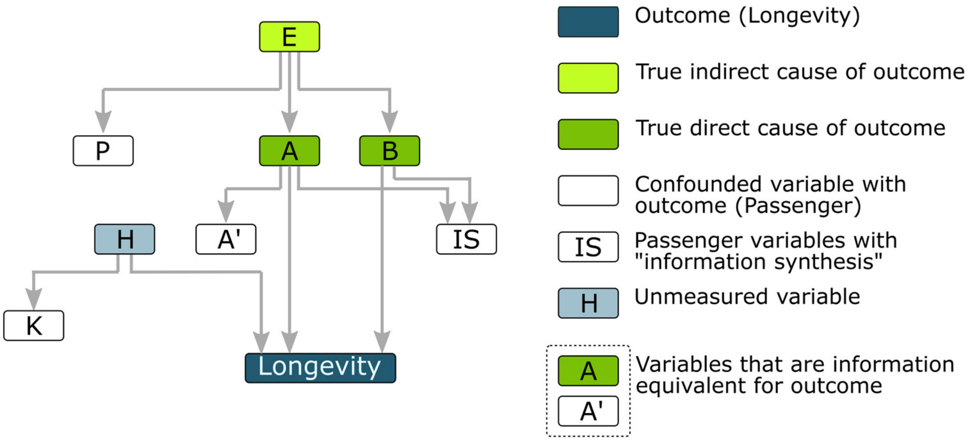
E
P
A
B
H
A'
IS
K
Longevity
Outcome (Longevity)
True indirect cause of outcome
True direct cause of outcome
Confounded variable with outcome (Passenger)
IS
Passenger variables with "information synthesis"
H
Unmeasured variable
A
A'
Variables that are information equivalent for outcome

Figure 1. A causal diagram illustrating the three main challenges in discovery of direct causes and Markov boundaries of longevity in simplified and idealized form.

Rectangles indicate variables, arrows indicate direct causal relationships, e.g., A is a direct cause of longevity. **Direct causes of longevity** are of interest because they have the **maximum combined causal effect on longevity** and are more amenable to practical discovery than the whole causal network. **Markov boundaries** are useful because they have maximal predictivity for longevity and maximum compactness.

Challenge 1=Feature selection, dimensionality reduction and classifier inductive biases. Mishandling measured confounding. (i) Non-causal feature selection often introduces unnecessary features resulting in unnecessarily large models. (ii) Such methods may also focus on non-causal variables that exhibit “information synthesis” (i.e., “signal aggregator” variables such as IS in the figure). Ranking by univariate association is such a commonly used method. (iii) Remote antecedent causes and related confounded variables (aka “passengers”) may also be preferentially selected (e.g., remote cause E and passenger P in the figure). (iv) It is also possible for various powerful predictor Machine Learning methods as well as classical statistical methods to be unable to differentiate between confounders and passengers and assign the same weights (e.g., Support Vector Machines and regularized regressors as well as Principal Component Analysis tend to view variables E and P, E and A, A and IS, as equally strong, despite the fact that they can be readily distinguished by conditional independence testing).

Challenge 2=Discovery methods are oblivious to equivalence classes or mishandling equivalences. Whenever a variable set has the exact same information about Longevity with another set, we say that they are Target Information Equivalent (TIE). For example, variables A and A’ are target information equivalent. This means that they have the same statistical information and characteristics with respect to Longevity. (i) Most analytic methods and protocols do not consider such equivalences and report a member of the class (e.g., either A or A’ in the example). Because the equivalence class can be vast (i.e., exponential to the number of variables in the dataset), true local causes can easily be ignored. The larger the equivalence class, the larger the probability that the true causes will be missed. (ii) Also, **collinearity analysis** is sometimes misunderstood to handle the problem, but this is not the case: (a) collinearities examine 2 variables at a time whereas information equivalency often also exists at the *variable set* level; (b) highly collinear variables may not be information equivalent for Longevity; (c) weakly collinear variables may be information equivalent for Longevity.

Challenge 3=Unmeasured Confounders. In the figure, H is an unmeasured confounder of K and Longevity. Algorithmic methods exist that under distributional assumptions, can reveal some of the unmeasured confounders or can ensure that some variables are not confounded by unmeasured variables. *However, no such methods exist in distributions with information equivalences* (as is the case in our study).

The analysis methodology employed in the present study outputs local causes of Longevity: {A, B, K} and its equivalent {A’, B, K}. These sets are all Markov boundaries thus have optimal predictive signal for Longevity and are minimal (i.e., maximally compact). They can be thus readily used to create optimal predictors. Except for unmeasured confounding of variable K, the methods used successfully avoid pitfalls and challenges outlined in Figure 1; of note, no non-experimental method exists to avoid this problem. The equivalency of A with A’ is identified and highlighted for further investigation. Passengers and information sinks plus remote causes are all filtered away. No measured direct causes are missed. Causal effects of all direct causes are correctly estimated. A relatively small number of experiments (up to the cardinality of the union of the local causal sets / Markov boundaries equivalence class) is needed to resolve both information equivalent sets *and* unmeasured confounding. Active learning algorithms exist to further reduce the number of experiments needed to resolve the equivalence class.

81], respectively, by additional independent hold-out validation (Table 1). For 10-year longevity, the best predictive performance was 0·66±0·02 and 0·66 [0·61,0·71] for cross-validation and independent hold-out, respectively, indicating that some of the baseline information could predict long-term longevity, even though predictivity was weaker. With 13 or fewer MB variables at each time horizon (Figure 3), predictivity was comparable to the complete set of 186 variables (Table 1). Although Levine Biological Age performed slightly better as a predictor of longevity than chronological age, both chronological age and Biological Age were weaker

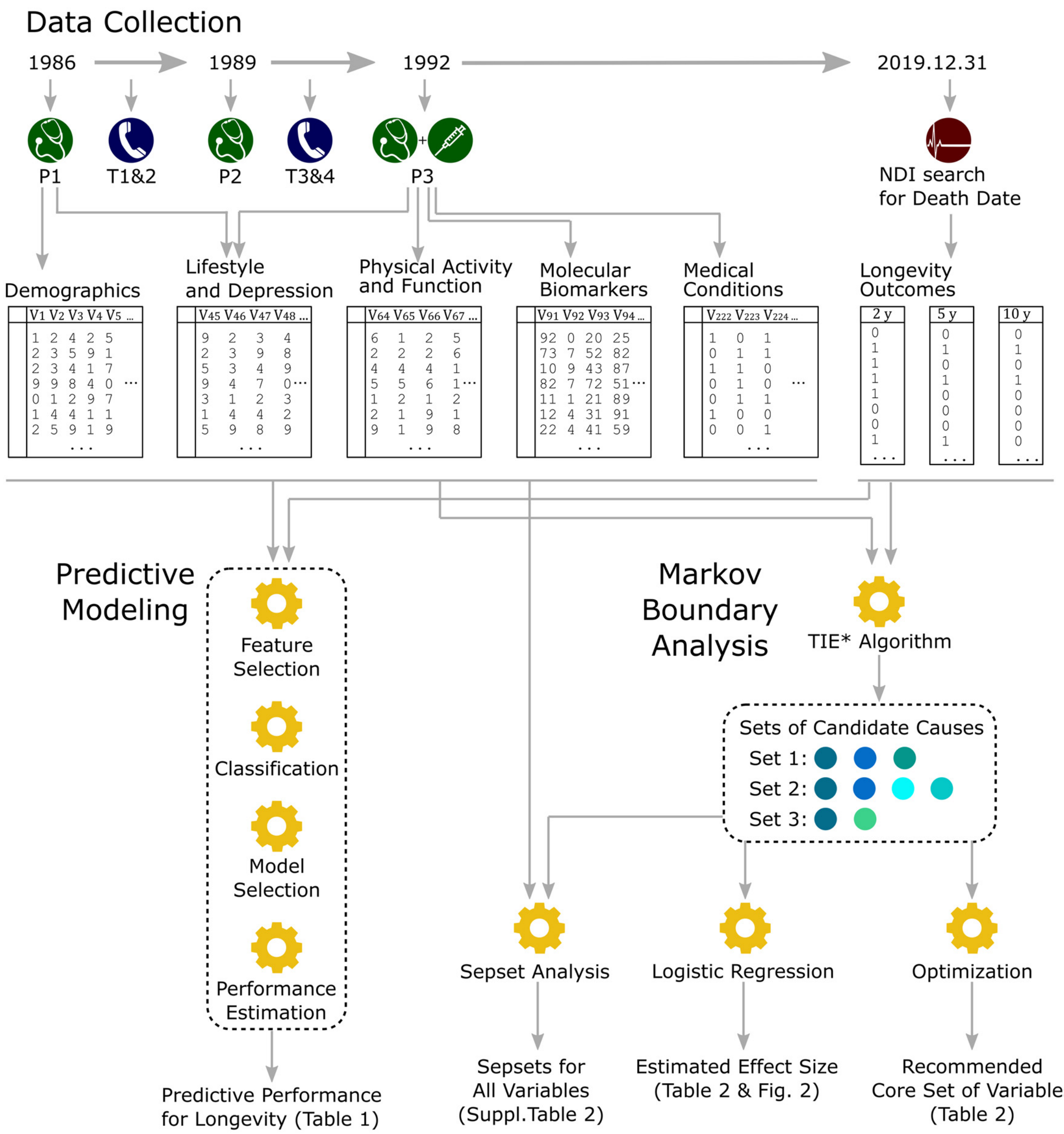

Figure 2. Schematic representation of D-EPESE study timeline and analyses.

The completed D-EPESE study was a longitudinal epidemiology assessment of locally representative community-dwelling older adults established in 1986, providing a wide range of measured clinical and molecular variables, and 33 years of subsequently documented survival data. Pertinent to this study, in-person (P1, P2 and P3) and telephone (T1&2, T3&4) interviews were conducted over 6 years from baseline in 1986. At P3, of those interviewed (2,569 survivors), 1727 provided consent for blood sampling and had a successful non-fasting blood draw; 1554 had blood stored for future use, and 1507 had both plasma and death data available for these analyses. Multiple (*N*=186) self-reported and clinically accessible measures (demographics, lifestyle and depression, physical activity and function, molecular biomarkers including clinical chemistries, haematological, lipids and metabolites by nuclear magnetic resonance spectroscopy (NMR), and medical conditions) were obtained. Mortality events were ascertained periodically by National Death Index searches; mortality was defined as death from any cause from 1992 (P3) through December 31, 2019 (27 years of follow-up after the blood was obtained) when the final National Death Index (NDI) search was performed. We modelled longevity at three different time horizons of clinical interest, each modelled separately, defined as a participant surviving more than or equal to 2 years, 5 years, and 10 years, respectively. The analytical pipeline included predictive modelling with a nested hold-out/cross-validation design, and Markov Boundary (MB, causal) analyses to identify local causes of longevity and their target information equivalent (TIE) classes ("signatures") at different time horizons. We also derived a sepset for each variable that was not part of any MB; sepset analysis elucidates which variables in a MB block the influence or subsume the information of a particular variable that is not part of the MB.

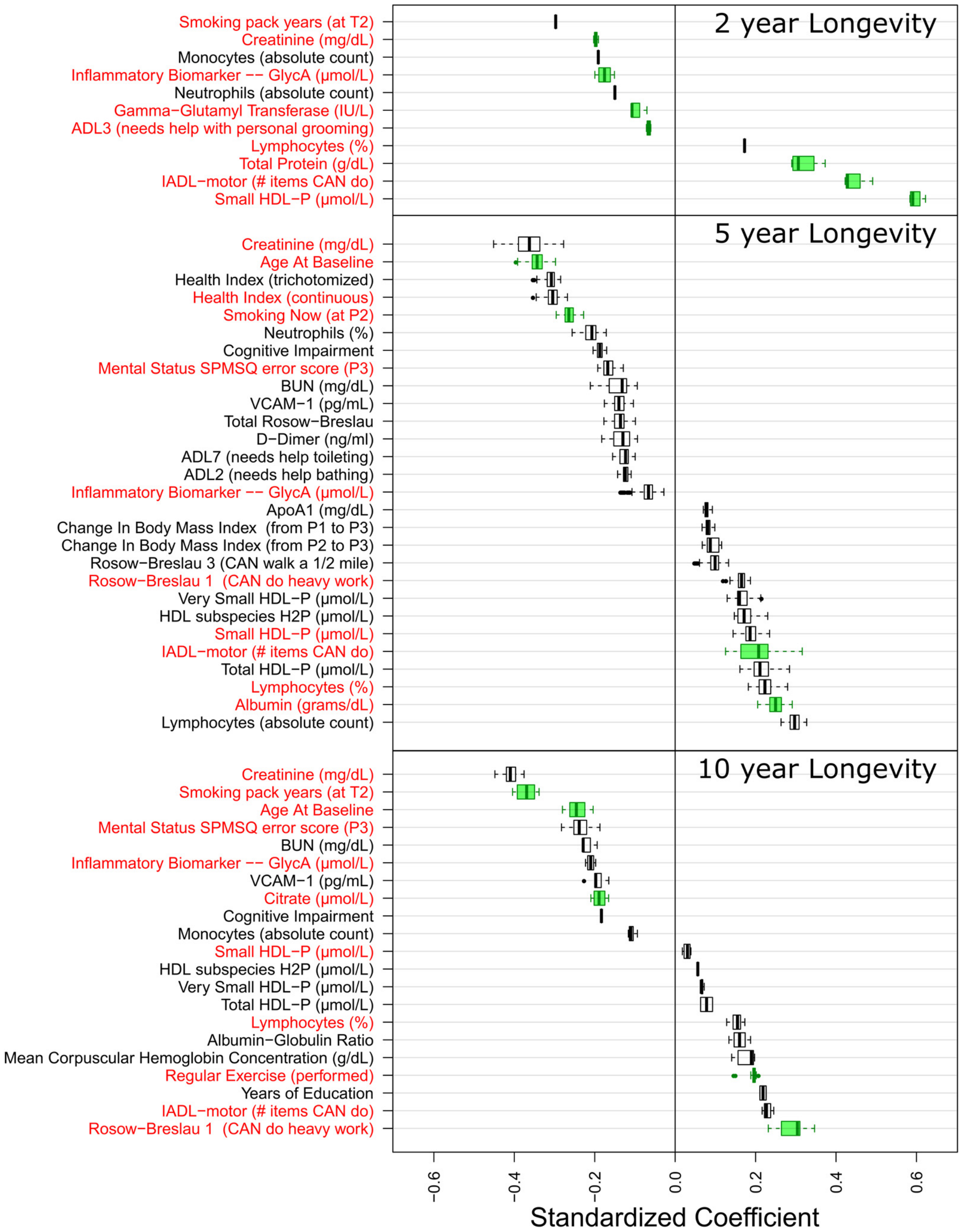

Figure 3. Variables that predict longevity.

Variables that predict 2-year (top panel), 5-year (middle panel) and 10-year (lower panel) longevity. The ranges of estimated effect sizes are depicted for variables in at least one Markov boundary (MB). All molecular variables were derived from measures at the time of the blood draw (third in-person evaluation, P3). Green indicates variables that appeared in all MBs. Total sample size 1507. The box in the figure represents the range of the estimates for each variable from models with the variable present. The left vertical line of the box represents the 1st quantile (Q1), the right line of the box represents the 3rd quantile (Q3), the internal line represents the median. left whisker=min(max(x), Q1 + 1.5 * IQR), right whisker=max(min(x), Q3−1.5 * IQR), where IQR is the inter-quartile range.

predictors and without added information compared to the EPESE variables (Table 1).

In addition, we examined time to death with a Cox model. Apart from mean corpuscular haemoglobin concentration (MCHC), the variables selected in the Cox model were identified as variables in MBs of one or more of the three time horizons. The performance estimation of the Cox model was 0·623+/-0·028 (C-index); the performance estimation of the hold-out testing set was 0·618.

Markov boundaries, local causes of longevity and their equivalence classes at different time horizons

Two-year. Four MBs predicted 2-year longevity and contained 8 variables, which in order of absolute standardized effect estimates were: small HDL particles (small HDL-P), the number of motor Instrumental Activities of Daily Living (IADL-motor) items able to be performed, total protein, GlycA, creatinine, gamma-glutamyl transferase (GGT), and Activity of Daily Living assessing level of help needed with personal grooming (ADL3) (Figure 3 and Supplementary Table 2a). Age was not among the local 2-year prediction variables, thus establishing that age is not a direct cause of 2-year longevity in the context of measured factors in our data. Three variables potentiated (small HDL-P, IADL-motor function (higher denoting better function) and total protein) and four attenuated (GlycA, creatinine, GGT, and more help needed with personal grooming) longevity. Given their appearance in all MBs, small HDL-P and motor function (IADL-motor) were the apparent direct causes of longevity among measured factors; they also had the greatest mean standardized effect estimates (95% CIs) from among the variables in the MBs-small HDL-P 0·60 [0·41, 0·78] and IADL-motor 0·44 [0·27, 0·58].

Five-year. 395 MBs were identified for 5-year longevity ranging in size from 10-15 (median 13) (Figure 3 and Supplementary Table 2b). Although the number of MBs increased dramatically compared to 2-year longevity, the total number of direct causes less than doubled. Four variables were in all MBs (largest to smallest absolute effect estimates): age, current cigarette smoking (at P2), albumin, and the number of IADL-motor items able to be performed. All MBs also contained one of the general health indices, an HDL particle measure, cognitive status, renal function (BUN or creatinine), and white blood cell (lymphocyte or neutrophil) percentage or absolute numbers, supporting their role as direct causes of 5-year longevity. More stable (less decline of) BMI prior to the blood draw was a potentiator of 5-year longevity. Consistent with the 2-year time horizon, lymphocyte measures (counts or percentage) were potentiators, whereas higher neutrophil measures were attenuators of longevity. Several circulating inflammation biomarkers were attenuators in multiple MBs, including VCAM-1, GlycA, and D-dimer. The majority of the predictive variables had similar absolute standardized effect estimates of ∼0·20 to 0·39. Age and current smoking were both attenuators and overall, the strongest predictors of 5-year longevity with mean standardized effect estimates of -0·34 and -0·26, respectively.

Ten-year. 138 equivalent MBs predicted 10-year longevity ranging in size from 8-11 (median 11) (Figure 3 and Supplementary Table 2c). Five variables were in all MBs, therefore directly causal among measured variables (absolute effect estimates from largest to smallest): pack years of cigarette smoking, ability to do heavy housework, age, regular exercise, and citrate. As for the 2- and 5-year time horizons, HDL particle variables were informative for prediction of 10-year longevity. However, HDL particle measures were selected in only half of the 10-year models; in most of the remaining models, IADL-motor function replaced HDL particle measures. Multiple markers were similarly predictive of 10-year longevity with absolute standardized effect estimates of ∼0·15 to 0·45. Smoking (attenuator) and ability to do heavy work (potentiator) variables were the strongest predictors of 10-year longevity with mean standardized effect estimates of -0·37 and 0·30, respectively.

Sepset analyses

At all three time horizons, self-reported function and small HDL particle variables were the most frequent sepsets; this means that these variables mediate or confound effects on longevity by most other variables that are not in MBs. At the 2-year time horizon, IADL-motor function was a sepset for age (Supplementary Table 3a), meaning that all information in age regarding longevity was contained in IADL-motor function. In univariable analyses, female sex was significantly associated [$P<0·05$ (t-test for continuous variables, Chi Square for categorical variables)] with greater 2-year (P = 0.001) and 5-year (P = 0.02), but not 10-year (P = 0.25) longevity. Although these results are consistent with the general observation that on average, women live longer than men, sex was not a direct cause of longevity since it was not part of any Markov boundaries. Nevertheless, among the 39 variables in Markov Boundaries, there were significant sex differences in 21 of them including age, education, smoking, SPMSQ (cognition), several physical function measures, total HDL-P, ApoA1, citrate, GlycA, MCHC, % and count of lymphocytes, count of monocytes, and creatinine.

Clinically recommended combined core set of variables

In a clinical care or research setting, a test to evaluate risk for 2-, 5-, and 10-year longevity simultaneously (rather than three separate tests for each time horizon) would be of potential value. With the goal of selecting a parsimonious core set (the smallest set but with full predictivity for all three time-horizons), we identified the variables of rows 8 and 10 for 2-year, rows 380 and 392 for 5-year, and row 25 for 10-year longevity (Supplementary Table 2). The estimated predictive performance of the core set for different time horizons is the same as for GLL models in Table 1. Our recommended core set of 17 variables (highlighted in red font in Figure 3 and Supplementary Tables 2a-c) include: 8 molecular biomarkers-small HDL-P, total protein, GlycA, creatinine, GGT, albumin, lymphocyte (percentage), and citrate; and 9 clinical biomarkers-IADL-motor, ADL3 (requires help with personal grooming), age, current cigarette smoking, pack year history of smoking, self-reported health index score (continuous), score on the Short Portable Mental Status Questionnaire (SPMSQ), Rosow-Breslau ability to do heavy housework (ROSOW BRES1), and regular exercise.

Discussion

We identified a relatively small number of putative direct causes of longevity from among 186 clinically accessible variables, with 8 to 15 variables containing the totality of signal for each time horizon. These variables had predictive performance AUCs of 0·76, 0·76, and 0·67 for predicting longevity at 2-, 5- and 10-years, respectively with 0.7≤AUC< 0.8 corresponding to acceptable discrimination per the Hosmer and Lemeshow interpretation system.[15] Greater concentrations of small HDL particles, younger age, and fewer pack years of cigarette smoking were the strongest determinants of longevity at 2-, 5- and 10-years, respectively. Interestingly, 13 indices used to build ePrognosis, a test that predicts risk of mortality from 6 months to 5 years for older adults in a variety of clinical settings, had C statistics of 0·70 or greater with a validated C statistic maximum of 0·79. Using medically accessible clinical and molecular data, our study achieved predictivity for the two- and five-year time horizons comparable to ePrognosis.[16] The nearer the time horizon, the stronger the predictivity of the models and the more molecular measures-as opposed to self-reported measures-were identified as predictors. The overall proportion of molecular, in contrast to clinical predictors of 2-year longevity, was 70%, supporting the concept that the baseline biomarkers best reflect the current biological mechanisms, known as endotypes, explaining the observable properties of the individual. Interestingly, race was not associated with longevity despite achieving good racial diversity (53% Black overall) of the sample; this is consistent with prior reports demonstrating the decreasing impact of race on life expectancy with aging with negligible differences beyond age 80.[17]

Physical function and HDL particle measures were consistently strong predictors of longevity at all time horizons. Age and cognitive function were direct causal factors of 5- and 10- but not 2-year longevity. Based on prediction of mortality over 5 or more years with 70-78 variables,[18, 19] the community-based Cardiovascular Health Study (CHS) identified some of the same factors as our study (such as physical activity, smoking and BMI) and indicated that objective, quantitative measures were better predictors of mortality than clinical history of disease (Fried 1998); however, the CHS analyses did not identify causal variables or their hierarchical relationship nor evaluate survival over shorter than the five year interval at which time horizon we identified factors that superseded age as causal for longevity.

As revealed by sepset analyses, for 2-year longevity, IADL-motor function was a more proximal causal predictor than 32 other variables, including age and other function measures. IADL-motor was present in 100% of MBs for 2- and 5-year longevity, indicating it is a direct cause for longevity at these time horizons. "Any regular exercise" was present in 100% of MBs for 10-year longevity. The presence of such modest physical activity in all MBs at all time horizons is concordant with studies showing that the beneficial health outcomes of physical activity begin when adopting even very modest amounts.[20-22] Our results support inferences of a causal nature that are further supported by bi-directional Mendelian randomization analyses, demonstrating that brisker walking pace is causally associated with longer leukocyte telomere length.[23] These results provide a compelling biological mechanism by which the longevity benefit of physical activity could be transduced.

The medically accessible *NMR LipoProfile* analysis yielded a strong lipoprotein predictor that appeared consistently across equivalent MBs, namely, the concentration of small HDL particles (<9 nm diameter), with greater concentrations potentiating longevity. Currently more than 215 different proteins are carried in HDL particles,[24] but their functions are not well-understood.[25] Consistent with the Cardiovascular Health Study of older adults that showed no association of HDL and total cholesterol with mortality,[18] it was notable in our study that total lipid concentrations (triglyceride, cholesterol) and cholesterol subfractions (including TRL-C, LDL-C and HDL-C), whether measured by NMR or by standard clinic-ordered lipid panel, were not among the most predictive and direct factors of longevity, contrary to numbers of small HDL particles that were highly predictive. The strong benefit of small HDL particles for longevity may be due to the well-known atheroprotective properties of HDL mediated by its ability to effect cholesterol efflux from peripheral tissues to the liver, as shown in *vitro* by cholesterol efflux from macrophages, especially in the presence of ABCA1 on the cell donating cholesterol to the particle.[26] *In vivo*, the atheroprotective properties of HDL also appear to depend on small HDL particle size and not HDL-cholesterol concentration.[27]

In addition, small but not large HDL particles bind, neutralize and clear endotoxins.[28] This intriguing finding supports a role of small HDL particles as a countermeasure to inflammation consistent with identification in this study of small HDL particles as positive predictors of longevity. Endotoxemia itself significantly decreases the number of small- and medium-sized HDL particles.[29] Thus, the role of higher number of small HDL particles as a promoter of longevity could be explained by atheroprotection, clearing endotoxin, and/or their indication of the lack of endotoxemia that would drive their clearance. These results suggest that NMR measures of small HDL particles are likely more informative than standard lipid panels for longevity prediction in a clinical setting.

Another NMR measure, GlycA, was in all MBs for 2-year and some MBs for 5- and 10-year longevity, suggesting a direct causal role for 2-year longevity and potentially for 5- and 10-year longevity. GlycA and smaller HDL subclasses had independent but opposite effects on mortality risk prediction.[30] GlycA, a marker of systemic inflammation, is a composite measure of glycan N-acetylglucosamine residues on enzymatically glycosylated acute-phase proteins,[31] including α1-acid glycoprotein, haptoglobin, α1-antitrypsin, α1-antichymotrypsin and transferrin.[32] To date, GlycA is positively associated with all-cause mortality in seven studies encompassing 63,180 individuals[33]; these studies reinforce our findings. In our study, cellular immune subsets (percent or count), including lymphocytes (potentiators), neutrophils and monocytes (attenuators), were also consistent predictors of longevity. Together with GlycA, these results underscore the importance of the immune system and inflammation status for longevity.

Key study strengths included: the clinical accessibility of all measures; the large number of measured variables; the long duration of follow-up; the large community-based sample; and the time when the samples were obtained that provided the opportunity to evaluate 'unadulterated' lipids prior to widespread use of statins (first available in 1987) that would make such a study more challenging today. Strengths of our analytic methodology that overcome limitations of prior studies, include: (a) the concurrent modelling of maximally predictive and maximally compact biomarker sets and signatures, and of causal factors, by using MB discovery; (b) the use of complete predictor and local causal discovery equivalence class algorithms producing *all putative local causal factor sets consistent with the data*; (c) identification of the biomarker set with maximal predictivity that is easiest to apply in clinical practice; and (d) sepset analysis that explains why some previously reported factors are rejected from our models specifically because their influence is subsumed by more proximal or confounding causal variables identified by our analysis.[34] In addition, the analysis protocol employed here had multiple procedures to prevent overfitting. Importantly, studies that report select variables within one or a few members of the optimal biomarker equivalence class (because the analytics employed are oblivious to the equivalences), will generate both false positives and negatives that our equivalence class modelling avoids.

There were also several study limitations. We expect our results to generalize to a population that is similar to our study population, however, due to selection bias, estimated causal effect estimations may not generalize to a different population, such as a population younger than 71 years of age. As mentioned previously, under the assumption of no unmeasured confounding variables, the Markov boundary members that were identified in all Markov boundary sets for a given outcome are the direct causes of the outcome and their causal effect estimates are unbiased. However, unmeasured confounding variables can lead to biased effect estimation. Although the number of variables examined in the current study greatly exceeds many prior studies, which leads to much reduced likelihood of encountering unmeasured confounding, unmeasured confounding may still exist from factors such as genetic variants, access to medical care and sufficient food, and level of chronic stress, among others. Future analyses with more variables will almost surely lead to discovery of even more proximal causes of longevity and elimination of false positive direct causes due to unmeasured confounders or mediators. Physical activities and exercise measurements were based on patient report and can be subjective; with the recent development of mobile health technology, incorporating activity tracking data could provide objective measurement for physical activity and might further improve the model. The D-EPESE biospecimens were a sample of convenience and not necessarily fasting; this variability may have limited the ability to detect an association of longevity with targeted metabolites but at the same time may have made the sample more representative of a clinic population. However, whether the samples were from fasting or non-fasting individuals should have no impact on the HDL particles or GlycA, although fasting conditions can impact the Triglyceride containing lipoproteins like TRL (VLDL). Those unable to give consent or provide blood were significantly more impaired (functionally and cognitively), and somewhat older; this could have imposed a ceiling effect on the relationships reported and limited the ability to identify a causal association of cognitive impairment with 2-year longevity.

In summary, longevity in older adults can be predicted to a considerable extent with a compact set of 8 to 15 readily accessible clinical variables with AUCs of 0·74, 0·77, and 0·67 for 2-, 5-, and 10-year longevity, respectively. Molecular measures are particularly dominant predictors of near-term (2-year) longevity. Age was a causal predictor of 5- and 10- but not 2-year longevity. HDL particle measures by NMR, physical function (both longevity potentiators), and GlycA (longevity attenuator) were selected in multiple models, suggesting their independent predictivity and causal relationship with longevity. A core set of 17 predictive variables may causally drive longevity or serve as proxies for more proximal but unmeasured determinants. Augmentation of clinical indices with this select set of medically accessible blood tests may benefit short-term longevity predictions. Small HDL particles may confer a longevity benefit by being both atheroprotective and endotoxin scavenging and thereby anti-inflammatory. The dominance of physical function over age in the near term, and causal association with longevity, even out to ten years, underscores the importance of even modest activity for prolonging life.

Contributors

VBK, SM, GGF, WEK, CFP, CFA conceived the study. SM, RT, GGF, BMB, DCP, MCO, CFP, CFA curated the data. SM, RT, BMB, DCP, CFP, CFA performed the formal statistical analyses. VBK, GGF, WEK, HJC, CFP, CFA acquired the funding for this study. VBK, GGF, MAC, JDP, HJC, CFA conducted the clinical and analytical investigations for this study. VBK, SM, RT, GGF, BMB, DCP, WEK, MAC, JDO, MCO, CFP, XZ, CFA devised the methodologies used for this study. MAC, JDO provided the NMR analyses. VBK, SM, GGF, WEK, CFA were responsible for all project administration. SM, RT, BMB and CFA were responsible for the computing resources, software, supervision, and data validation. VBK, GGF, BMB, DCP, WEK, HJC, CFP provided study material resources. VBK was responsible for the overall project supervision. VBK, SM, RT, CFA produced the graphics for visualization of the results. VBK, SM, CFA drafted the manuscript. All authors confirm that they had full access to all the data in the study and accept responsibility to submit for publication. VBK, SM, CFP and CFA had direct access and verify the underlying data reported in the manuscript. All authors read and approved the final version of the manuscript.

Data sharing statement

Deidentified participant data related to the EPESE cohort and codes that pertain to methods new to this paper (i.e., not previously published) are available upon request from the corresponding author with publication under a data sharing agreement.

Declaration of interests

Drs. Connelly and Otvos are employees of and own stock in Labcorp, the commercial provider of the NMR LipoProfile blood test. Additional institutional NIH funding is declared for Dr. Zhang (RO1 AG070146) and Dr. Ma (RO1AG070146 and RO1 HL153497) and consulting fees to Dr. Ma related to this work from the Duke Claude D. Pepper Older Americans Independence Center NIH/NIA P30-AG028716 grant. The remaining authors declare no competing interests. The funding sources provided funding only and had no role in writing of the manuscript or the decision to submit it for publication. No author has been paid to produce this manuscript. The authors were not precluded from accessing data in the study, and they accept responsibility to submit for publication.

Acknowledgements

This was work was supported by NIH/NIA R01AG054840 (to VBK, SM, RT, GGF, BMB, DCP, WEK, HJC, MCO, CFP, and CFA), R01AG12765 (to GGF, and CFP), the Duke Claude D. Pepper Older Americans Independence Center NIH/NIA P30-AG028716 (to VBK, GGF, WEK, HJC, and CFP), NIH/NIA Contract N01-AG12102 (to GGF), and NIH/NCATS UL1TR002494 (to SM, RT, and CFA).

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2022.104292.

References

1 Nelson PG, Promislow DEL, Masel J. Biomarkers for aging identified in cross-sectional studies tend to be non-causative. *J Gerontol Series A, Biol Sci Med Sci*. 2020;75(3):466–472.

2 Social Security Administration. Actuarial life table. 2017. p. https://www.ssa.gov/OACT/STATS/table4c6.html.

3 Yourman LC, Lee SJ, Schonberg MA, Widera EW, Smith AK. Prognostic indices for older adults: a systematic review. *JAMA*. 2012;307(2):182–192.

4 Cornoni-Huntley J, Blazer D, Lafferty M, Everett D, Brock D, Farmer M. *Established Populations for Epidemiologic Studies of the Elderly: Resource Data Book*. Washington DC: PHS, NIH; 1990.

5 Huffman KM, Pieper CF, Kraus VB, Kraus WE, Fillenbaum GG, Cohen HJ. Relations of a marker of endothelial activation (s-VCAM) to function and mortality in community-dwelling older adults. *J Gerontol A Biol Sci Med Sci*. 2011;66(12):1369–1375.

6 Jeyarajah EJ, Cromwell WC, Otvos JD. Lipoprotein particle analysis by nuclear magnetic resonance spectroscopy. *Clin Lab Med*. 2006;26(4):847–870.

7 Matyus SP, Braun PJ, Wolak-Dinsmore J, et al. NMR measurement of LDL particle number using the Vantera Clinical Analyzer. *Clin Biochem*. 2014;47(16-17):203–210.

8 Connelly MA, Wolak-Dinsmore J, Dullaart RPF. Branched chain amino acids are associated with insulin resistance independent of leptin and adiponectin in subjects with varying degrees of glucose tolerance. *Metab Syndr Relat Disord*. 2017;15(4):183–186.

9 Parker DC, Bartlett BN, Cohen HJ, et al. Association of blood chemistry quantifications of biological aging with disability and mortality in older adults. *J Gerontol A Biol Sci Med Sci*. 2020;75(9):1671–1679.

10 Statnikov A, Lytkin NI, Lemeire J, Aliferis CF. Algorithms for discovery of multiple Markov boundaries. *J Mach Learn Res*. 2013;14:499–566.

11 Aliferis C, Statnikov A, Tsamardinos I, Mani S, Koutsoukos X. Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: analysis and extensions. *J Mach Learn Res*. 2010;11:253–284.

12 Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd Ed. Cambridge: Cambridge University Press; 2009.

13 Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. 2nd Ed. MIT Press; 2001:568.

14 Cornoni-Huntley J, Blazer D, Lafferty M, Everett D, Brock D, Farmer M. Established populations for epidemiologic studies of the elderly. *Resource Data Book. Vol II*. Washington DC: PHS, NIH; 1990.

15 Hosmer D, Lemeshow S. *Model-Building Strategies and Methods for Logistic Regression: Applied Logistic Regression*. 2nd Ed. New York: John Wiley and Sons; 2000.

16 ePrognosis [Internet]. UCSF. 2011 [cited January 4, 2022]. Available from: https://eprognosis.ucsf.edu/.php.

17 Hummer R, Benjamins M, Rogers R. Racial and ethnic disparities in health and mortality among the U.S. elderly population. In: Anderson N, Bulatao R, Cohen B, eds. *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*. Washington, DC: National Academies Press; 2004:53–94.

18 Fried LP, Kronmal RA, Newman AB, et al. Risk factors for 5-year mortality in older adults: the cardiovascular health study. *JAMA*. 1998;279(8):585–592.

19 Newman AB, Sachs MC, Arnold AM, et al. Total and cause-specific mortality in the cardiovascular health study. *J Gerontol Series A, Biol Sci Med Sci*. 2009;64(12):1251–1261.

20 Saint-Maurice PF, Troiano RP, Berrigan D, Kraus WE, Matthews CE. Volume of light versus moderate-to-vigorous physical activity: similar benefits for all-cause mortality? *J Am Heart Assoc*. 2018;7 (7).

21 Kraus WE, Powell KE, Haskell WL, et al. Physical activity, all-cause and cardiovascular mortality, andcCardiovascular disease. *Med Sci Sports Exerc*. 2019;51(6):1270–1281.

22 Kraus VB, Sprow K, Powell KE, et al. Effects of physical activity in knee and hip osteoarthritis: a systematic umbrella review. *Med Sci Sports Exerc*. 2019;51(6):1324–1339.

23 Dempsey PC, Musicha C, Rowlands AV, et al. Investigation of a UK biobank cohort reveals causal associations of self-reported walking pace with telomere length. *Commun Biol*. 2022;5(1):381.

24 Vickers KC, Michell DL. HDL-small RNA export, transport, and functional delivery in atherosclerosis. *Curr Atheroscler Rep*. 2021;23(7):38.

25 Mutharasan RK, Thaxton CS, Berry J, et al. HDL efflux capacity, HDL particle size, and high-risk carotid atherosclerosis in a cohort of asymptomatic older adults: the Chicago Healthy Aging Study. *J Lipid Res*. 2017;58(3):600–606.

26 Du XM, Kim MJ, Hou L, et al. HDL particle size is a critical determinant of ABCA1-mediated macrophage cellular cholesterol export. *Circ Res*. 2015;116(7):1133–1142.

27 McGarrah RW, Craig DM, Haynes C, Dowdy ZE, Shah SH, Kraus WE. High-density lipoprotein subclass measurements improve mortality risk prediction, discrimination and reclassification in a cardiac catheterization cohort. *Atherosclerosis*. 2016;246:229–235.

28 Määttä AM, Salminen A, Pietiäinen M, et al. Endotoxemia is associated with an adverse metabolic profile. *Innate Immun*. 2021;27(1):3–14.

29 Pirillo A, Catapano AL. Norata GD. HDL in infectious diseases and sepsis. *Handb Exp Pharmacol*. 2015;224:483–508.

30 McGarrah RW, Kelly JP, Craig DM, et al. A novel protein glycan-derived inflammation biomarker independently predicts cardiovascular disease and modifies the association of HDL subclasses with mortality. *Clin Chem*. 2017;63(1):288–296.

31 Connelly MA, Otvos JD, Shalaurova I, Playford MP, Mehta NN. GlycA, a novel biomarker of systemic inflammation and cardiovascular disease risk. *J Transl Med*. 2017;15(1):219.

32 Otvos JD, Shalaurova I, Wolak-Dinsmore J, et al. GlycA: a composite nuclear magnetic resonance biomarker of systemic inflammation. *Clin Chem*. 2015;61(5):714–723.

33 Gruppen EG, Kunutsor SK, Kieneker LM, et al. GlycA, a novel pro-inflammatory glycoprotein biomarker is associated with mortality: results from the PREVEND study and meta-analysis. *J Intern Med*. 2019;286(5):596–609.

34 Spirtes P, Glymour C, Scheines R. *Causation, Prediction, and Search*. 2nd Ed. MIT Press; 2001:568 p.