Predicting mortality from 57 economic, behavioral, social, and psychological factors

Eli Puterman[a,1], Jordan Weiss[b,2], Benjamin A. Hives[a], Alison Gemmill[c], Deborah Karasek[d], Wendy Berry Mendes[e], and David H. Rehkopf[f,1]

[a]School of Kinesiology, The University of British Columbia, Vancouver, BC V6T1Z1, Canada; [b]Population Studies Center and the Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA 19104; [c]Department of Population, Family and Reproductive Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [d]Preterm Birth Initiative, Department of Obstetrics, Genecology & Reproductive Sciences, University of California, San Francisco, CA 94158; [e]Department of Psychiatry, University of California, San Francisco, CA 94143; and [f]School of Medicine, Stanford University, Palo Alto, CA 94304

Edited by Eileen M. Crimmins, University of Southern California, Los Angeles, CA, and approved May 11, 2020 (received for review October 22, 2019)

Behavioral and social scientists have identified many nonbiological predictors of mortality. An important limitation of much of this research, however, is that risk factors are not studied in comparison with one another or from across different fields of research. It therefore remains unclear which factors should be prioritized for interventions and policy to reduce mortality risk. In the current investigation, we compare 57 factors within a multidisciplinary framework. These include (*i*) adverse socioeconomic and psychosocial experiences during childhood and (*ii*) socioeconomic conditions, (*iii*) health behaviors, (*iv*) social connections, (*v*) psychological characteristics, and (*vi*) adverse experiences during adulthood. The current prospective cohort investigation with 13,611 adults from 52 to 104 y of age (mean age 69.3 y) from the nationally representative Health and Retirement Study used weighted traditional (i.e., multivariate Cox regressions) and machine-learning (i.e., lasso, random forest analysis) statistical approaches to identify the leading predictors of mortality over 6 y of follow-up time. We demonstrate that, in addition to the well-established behavioral risk factors of smoking, alcohol abuse, and lack of physical activity, economic (e.g., recent financial difficulties, unemployment history), social (e.g., childhood adversity, divorce history), and psychological (e.g., negative affectivity) factors were also among the strongest predictors of mortality among older American adults. The strength of these predictors should be used to guide future transdisciplinary investigations and intervention studies across the fields of epidemiology, psychology, sociology, economics, and medicine to understand how changes in these factors alter individual mortality risk.

mortality | transdisciplinary | social | behavioral | data-driven

Population aging has given rise to an increase in noncommunicable diseases that account for the majority of deaths in the United States (1). Although disease and mortality are caused by molecular, cellular, and physiological changes, nonbiological processes play important roles in shaping mortality risk. The National Institutes of Health Office of Behavioral and Social Sciences designates the behavioral and social sciences as fundamental to our understanding of disease pathogenesis and mortality, with knowledge derived from these disciplines leading to scientific breakthroughs that can transform health policy (2). New efforts and approaches to the ways in which research guides policy are needed given the three-decade stagnation of US life expectancy relative to other industrialized countries (3).

Studies of the determinants of mortality have identified a wide range of behavioral risk factors across disciplines. McGinnis and Foege (4), followed by the work of Mokdad and colleagues (5), established the prevailing role that health behaviors—predominantly smoking, poor nutrition, and physical inactivity—play in mortality rates, accounting for nearly 35% of all deaths in the United States. While health behaviors have received the majority of attention in the scientific literature, other studies have identified specific economic, social, and psychological factors associated with higher risk of mortality, such as early childhood adversity (6), financial difficulties in adulthood (7), poor social relationships (8), lower levels of neuroticism and conscientious (9, 10), and experiences of discrimination (11). An important limitation of much of this research, however, is that risk factors within and between these domains are often studied in isolation from each other with a priori hypotheses, so it is unclear which factors are the relatively strongest predictors of mortality risk. Additionally, a focus on single specific risk factors can result in publication of inflated effect sizes (12).

Several studies on mortality have sought to move beyond these siloed single-factor hypothesis, testing approaches to incorporate independent predictors from across domains. For example, Ganna and Ingelsson (13) investigated 655 health, demographic, and lifestyle predictors of 5-y mortality in nearly 500,000 adults in the United Kingdom. These risk factors included circulating blood biomarkers, anthropometrics, health and medical histories, sociodemographics, early life health factors, family history, psychosocial factors, and health behaviors. Following an examination of the 655 factors in independent analyses, proxy measures of health itself, as measured through self-reported health, recent morbidities, disability, medication use, and walking pace, were, not surprisingly, the strongest predictors of all-cause mortality, as they capture the most proximal underlying pathophysiology preceding death. In contrast, when those with recent illnesses were excluded, smoking emerged as an important predictor of mortality.

Ganna and Ingelsson's investigation highlights the importance of an approach that incorporates factors from across different

Significance

In our prospective study using nationally representative data from 13,611 adults in the US Health and Retirement Study, we used traditional and machine-learning statistical approaches to reveal the most important factors across the behavioral and social sciences that predict mortality in older adults. In the study, we found that top predictors of mortality spanned all investigated domains, opening up opportunities for future hypothesis generation in observational and clinical studies and the identification of potential new targets for screening and policy.

Author contributions: E.P., W.B.M., and D.H.R. designed research; E.P., J.W., and D.H.R. performed research; E.P., J.W., B.A.H., A.G., D.K., and D.H.R. analyzed data; and E.P., J.W., B.A.H., A.G., D.K., W.B.M., and D.H.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: eli.puterman@ubc.ca or drehkopf@stanford.edu.

[2]Present address: Department of Demography, University of California, Berkeley, CA 94720.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1918455117/-/DCSupplemental.

First published June 22, 2020.

domains to determine the strongest predictors of mortality. However, the inclusion of proximal health factors obstructs the discovery of behavioral and social factors from across the lifespan. A simultaneous investigation focused on predictors in the economic, behavioral, social, and psychological domains from across the life course will help advance our understanding of the importance of these factors as earlier, longer-term predictors of mortality (14). Such an investigation may potentially reveal important transdisciplinary life course contributors to mortality to inform future investigations, similar to genome-wide association studies (GWASs) (15, 16) and environment-wide association studies (EWASs) (17–19) that have revealed specific genetic factors or environmental toxins that are predictive of longevity, ill health, and mortality. Importantly, restricting our analyses to behavioral and social factors from across the life course, including childhood, establishes the importance of factors within these domains to include in future studies of the exposome (20) or in studies, such as that of Ganna and Ingelsson, which traverse the behavioral and social with the biological and medical sciences.

Comprehensively ranking independent predictors of mortality from across the behavioral and social sciences requires unique data that include the most relevant potential risk factors. We used data from the US Health and Retirement Study (HRS). The HRS is a nationally representative study of adults aged 50 y and older, with biannual data collection since its inception in 1992. Adults included in the current investigation were aged 52 to 104 y at the time the exposures were measured between the years of 1992 and 2008, with 6 y of follow-up for mortality. A total of 13,611 adults are included in the current analysis.

The HRS is not only unique in its comprehensive investigation of socioeconomic conditions, but also in its inclusion of several factors from the behavioral and social sciences, including early life factors, social relationships, and psychological characteristics. In the current investigation, we use three complementary analytical approaches to determine the relative contribution to prospective mortality of 57 independent predictors measured from years 1992 to 2008 selected from six commonly investigated domains in the behavioral and social sciences: (*i*) adverse socioeconomic and psychosocial experiences during childhood and (*ii*) socioeconomic conditions, (*iii*) health behaviors, (*vi*) social connections, (*v*) psychological characteristics, and (*vi*) adverse experiences during adulthood. We further explore how each of these 6 identified domains, and the combination across all 57 factors, predict mortality.

First, we examine the contribution of each individual predictor, regardless of domain, to time to death using Cox regression models. Within the Cox regression framework, we also explore the proportion of the variance explained by each domain alone to mortality and the proportion explained by all 57 factors when included in the same analysis. Our analysis also capitalizes on advances in statistical science that have paved the way for data-driven approaches that simultaneously analyze a large number of predictors and their interactions (2, 21). Our analyses include these data-driven approaches, namely the least absolute shrinkage and selection operator (lasso) regression (22) and random forest survival analysis (RFSA) (23). Both of these approaches have contributed to understanding novel interdisciplinary and interactive pathways to health and mortality while minimizing prediction error, such as over- or underestimating the predictive power of each factor, that can occur with traditional Cox regression models or survival decision trees when using data from individual datasets. Finally, we completed a replication of our primary Cox regression analyses in an independent dataset using another United States cohort of participants, the Midlife in the United States (MIDUS) Study. MIDUS offers the most comprehensive set of variables in the United States that best match the social and behavioral factors considered here in our primary study population (HRS).

Results

Participants. The average age of participants was 69.3 y (SD = 9.7), with the majority women (58.6%), white (77.6%), and born in the United States (91.0%).

Descriptive Statistics. *SI Appendix*, Table S1, presents descriptive statistics for all 57 variables included in the study.

Main Results.

Estimation of mortality risk for each independent factor. We used Cox regression to estimate the excess mortality risk of having a particular level of exposure to a risk factor compared to no exposure in 57 independent models adjusted for the following demographic characteristics (corresponding hazard ratio): male gender (HR = 1.28, 95% CI = 1.23, 1.33), race/ethnicity (HR_{Black} = 1.22, 95% CI = 1.13, 1.31; $HR_{Other\ Race}$ = 1.08, 95% CI = 0.94, 1.24; $HR_{Hispanic}$ = 1.02, 95% CI = 0.92, 1.12), and whether the individual was foreign-born (HR = 0.87, 95% CI = 0.80, 0.95). Table 1 and Fig. 1 show the individual hazard ratios and confidence intervals for each predictor ranked from strongest to weakest association with mortality over the study period (2008 to 2014). Point estimates for the hazard ratios are represented by dots; the line widths for each predictor present the 95% confidence intervals adjusted for multiple comparisons. Confidence intervals in our study that do not include 1 are considered statistically significant at the 95% level.

The 10 factors associated with the greatest risk of mortality over the study period were current or previous history as a smoker (HR = 1.91, 95% CI = 1.70, 2.14 and HR = 1.32, 95% CI = 1.22, 1.43, respectively), history of divorce (HR = 1.44, 95% CI = 1.31, 1.60), history of alcohol abuse (HR = 1.36, 95% CI = 1.14, 1.61), recent financial difficulties (HR = 1.32, 95% CI = 1.22, 1.43), history of unemployment (HR = 1.32, 95% CI = 1.10, 1.59), lower life satisfaction (HR = 1.31, 95% CI = 1.19, 1.45), never married (HR = 1.30, 95% CI = 1.03, 1.63), history of food stamps (HR = 1.28, 95% CI = 1.09, 1.49), and negative affectivity (HR = 1.23, 95% CI = 1.14, 1.33). Of the 57 predictors, 42 had confidence intervals that did not include 1, substantiating many previous a priori studies on these individual factors.

In order to estimate heterogeneity of associations by race, gender, education, or age group, we completed a series of follow-up analyses with interactions between each of these factors with each of the 57 factors considered here. We found little evidence that the strength of prediction differed by gender (*SI Appendix*, Table S2), between white and nonwhite participants (*SI Appendix*, Table S3), or between those who completed high school and those who did not (*SI Appendix*, Table S4). Many significant associations were apparent for individuals younger than 75 y old, whereas, for those 75 y and older, significant associations were sparse (*SI Appendix*, Table S5). Results were consistent when we censored participants who died within 2 y of 2008, where illness prior to death may be more likely to affect the level of the exposure (*SI Appendix*, Table S6).

Estimation of mortality risk for each domain and all domains combined. Next, we calculated the proportion of variance explained from Cox regression models including (*i*) all demographic characteristics combined, (*ii*) demographic characteristics and specific domains, and (*iii*) one final model with all 57 variables. Estimates are conditional on age as the baseline hazard, but proportion of variance explained does not include age. Demographic characteristics of gender, race/ethnicity, and place of birth explained 1.9% of the variation in mortality risk over the study period, and each domain independently provided an additional 1.1 to 4.7% increase in predictive power, with the smallest difference for childhood economic and psychosocial adverse experiences and largest difference for health

Table 1. Hazard ratios and 95% confidence intervals for each predictor ranked from strongest to weakest association with mortality over the study period

Variable	HR	95% CI
Current smoker	1.91	1.70, 2.14
History of divorce	1.45	1.31, 1.60
Alcohol abuse	1.36	1.15, 1.61
Recent financial difficulties	1.32	1.22, 1.44
History of unemployment	1.32	1.10, 1.59
History of smoking	1.32	1.22, 1.43
Lower life satisfaction	1.31	1.19, 1.45
Never married	1.30	1.03, 1.63
History of food stamps	1.28	1.09, 1.49
Negative affectivity	1.23	1.14, 1.33
Negative interactions with family	1.23	1.12, 1.35
Negative interactions with children	1.22	1.12, 1.34
Daily discrimination	1.22	1.12, 1.32
Trait anxiety	1.21	1.12, 1.31
Lower positive interactions with children	1.21	1.11, 1.31
Childhood psychosocial adversities	1.20	1.11, 1.31
Anger out	1.18	1.08, 1.28
Major discrimination	1.17	1.07, 1.29
Negative Interactions with friends	1.17	1.08, 1.27
Cynical hostility	1.16	1.07, 1.26
Pessimism	1.16	1.07, 1.25
Low/no vigorous activity	1.15	1.03, 1.28
History of food insecurity	1.14	1.02, 1.27
Hopelessness	1.14	1.06, 1.23
Lower positive affectivity	1.14	1.05, 1.23
Lower optimism	1.13	1.05, 1.22
Lower occupational status	1.13	1.03, 1.25
Lower wealth	1.13	0.98, 1.30
History of Medicaid	1.13	1.01, 1.26
Lower neighborhood safety	1.13	1.04, 1.23
Anger in	1.13	1.04, 1.22
Loneliness	1.12	1.04, 1.21
Lower neighborhood cohesion	1.12	1.04, 1.20
Low/no moderate activity	1.12	1.04, 1.20
Lower purpose in life	1.12	1.03, 1.21
Neighborhood disorder	1.11	1.03, 1.20
Sleep problems	1.11	1.02, 1.20
Lower positive interactions with family	1.11	1.03, 1.20
Lower conscientiousness	1.10	1.03, 1.17
Perceptions of obstacles	1.09	1.02, 1.17
Lower neuroticism	1.09	1.02, 1.16
Lower extroversion	1.09	1.01, 1.17
Lower income	1.07	0.88, 1.30
Lower education	1.07	0.99, 1.16
Family received financial help in childhood	1.07	0.96, 1.19
Lower sense of mastery	1.06	0.99, 1.14
Lower father occupational status	1.06	0.98, 1.14
History of renting	1.04	0.96, 1.14
Relocated homes in childhood	1.03	0.95, 1.13
Lower openness to experiences	1.02	0.95, 1.10
Lower religiosity	1.00	0.92, 1.08
Lower education father	0.98	0.87, 1.10
Father was unemployed in childhood	0.97	0.89, 1.05
Lower positive interactions with friends	0.96	0.89, 1.04
Lower agreeableness	0.94	0.88, 1.02
Adulthood psychosocial adversity	0.93	0.86, 1.01
Lower education mother	0.86	0.77, 0.97

behaviors. Social connections, such as objective markers of marital status and subjective reports of positive and negative relationships with family and friends, ranked second in the increase in predictive power for mortality (3.1%). These were followed by psychological characteristics (2.6% increase), adulthood socioeconomic conditions (2.3% increase), and adulthood adverse experiences (1.3% increase). All variables together predicted an additional 9.5% of the variance in

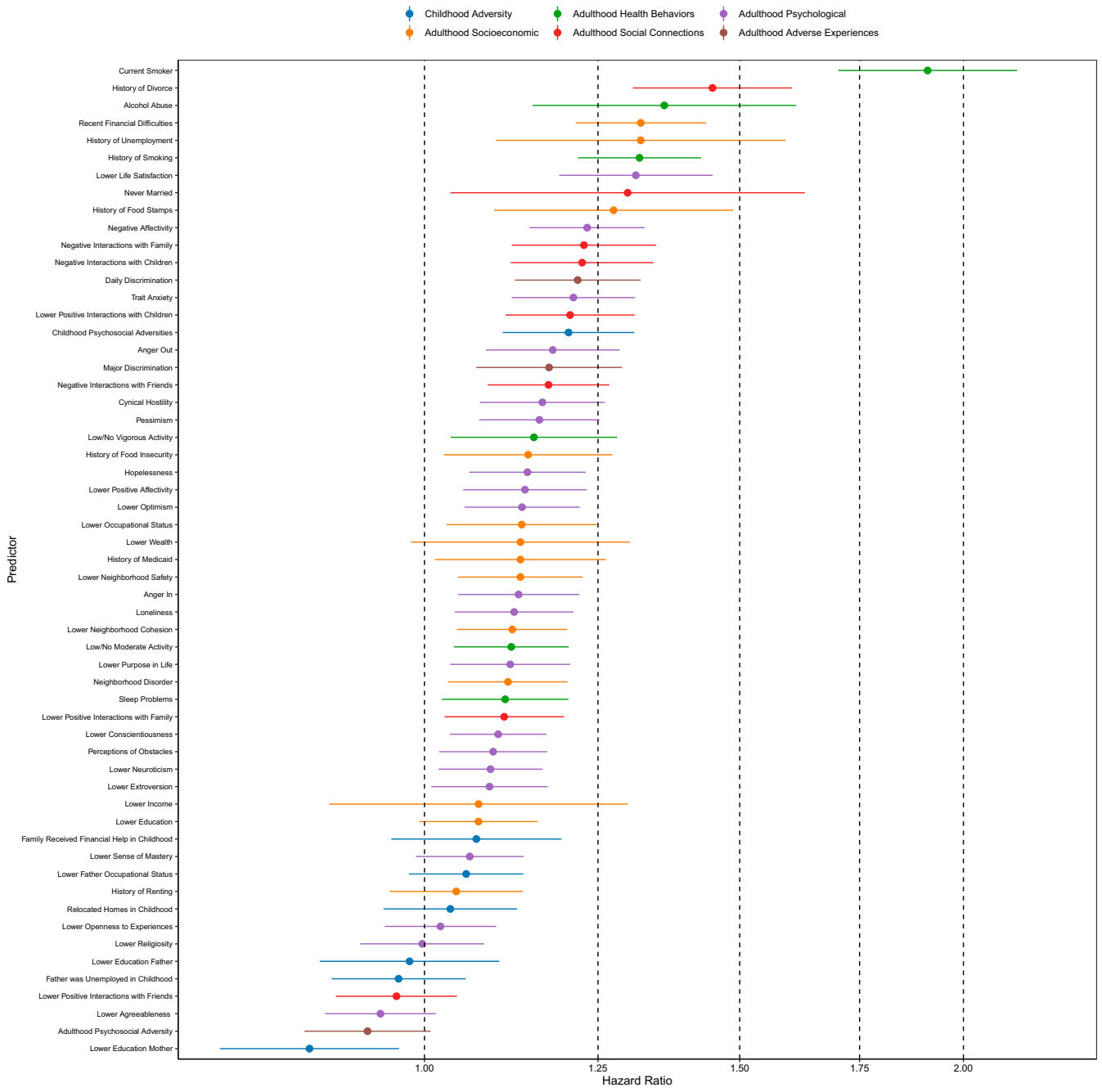

Fig. 1. Independent Cox regression hazard ratios of each predictor for mortality. Confidence intervals that include 1 indicate that a predictor is not statistically significant at the 5% level, corrected for multiple tests using the Bonferroni method (24). Age is used as the baseline hazard. Larger hazard ratios indicate higher mortality risk.

mortality in older adults above that of demographic characteristics, suggesting mostly unique variance in mortality accounted for by each domain.

As each domain included a different number of variables, we also ran principal component analyses within each domain and selected the top three principal components within each domain to predict mortality. In this approach, each domain predicted an additional 0.7 to 4.4% of variance, maintaining the same rank of lowest to greatest proportion of variance explained. Specifically, childhood factors ranked lowest (0.7% increase), followed by adulthood adverse experiences (1.2% increase), adulthood socioeconomic conditions (1.3% increase), psychological characteristics (1.3% increase), social connections (1.9% increase), and finally behavioral factors (4.4% increase), which ranked highest. Combined across all domains, all principal components explained an additional 7.2% of the variance in mortality.

Estimation of mortality using lasso and random forest survival analysis. For a second complementary approach, we implemented a regression algorithm that allows a large number of predictors to be fit in a single model: lasso regression (22). As an alternative to prediction with several variables at a time, lasso allows an analysis of all 57 risk factors and demographic factors simultaneously. In order to produce robust estimates, the algorithm estimates hazard ratios that are shrunk closer to the null, with

cross-validation used to produce the best-fitting parsimonious model. We present in Fig. 2 hazard ratios estimated from lasso survival analysis (22). The 10 strongest behavioral and social predictors of mortality were largely similar to the results from the independent Cox survival models, even though all factors were included in the same model. The additional top 10 most predictive factors identified with the lasso model were negative interactions with children, which was also a strong predictor in the Cox survival models, and childhood psychosocial adversity.

Finally, we used a random forest algorithm to determine if there were any additional factors identified through this approach, which additionally demonstrated that household wealth was a leading predictor of mortality (*SI Appendix*, Fig. S1).

Replication in the MIDUS study. Thirty-nine of the 57 variables across the 6 domains were able to be matched between the MIDUS Study and the HRS (*SI Appendix*, Table S7). The replication of the unweighted Cox regression models in the MIDUS study sample demonstrated generally similar unweighted hazard ratios and overlapping confidence intervals across 85% of factors from the HRS (*SI Appendix*, Table S8), without accounting for any corrections for multiple comparisons. Additionally, for the 10 highest ranked predictors in the HRS in this refined set of 39

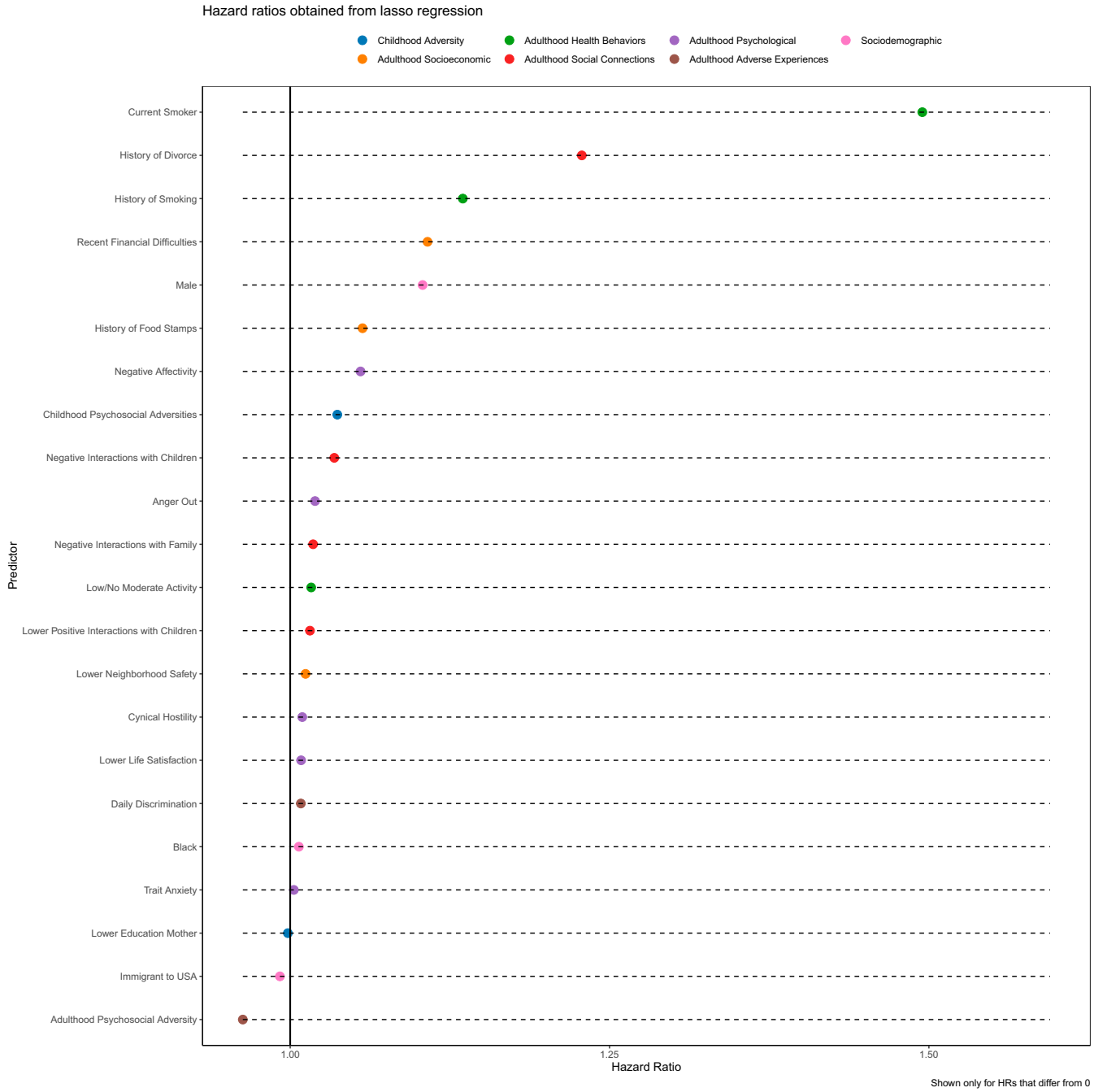

Fig. 2. Hazard ratios from time-to-event lasso regression. Age is used as the baseline hazard. All 57 factors were used to fit the model simultaneously, along with demographic factors of male gender, black race, other race, and Hispanic ethnicity. Only hazard ratios that differ from 0 in the final model are presented in the figure.

variables, 5 directly matched the 10 highest ranked in MIDUS, including past or current smoking status, alcohol abuse, history of employment, and negative affectivity. Two additional factors in the HRS's highest ranked, life satisfaction and recent financial difficulties, are closely linked conceptually to hopelessness and wealth, which were among MIDUS's highest ranked. There were some exceptions to the overlapping confidence intervals, however, with a lack of overlap in confidence intervals in 15% ($n = 6$) of the 39 factors, including history of divorce, history of renting, smoking status, lower positive interactions with friends, hopelessness, and lower neuroticism (Fig. 3 shows HRs and confidence intervals from both studies). Of these, the hazard ratios for smoking, divorce, and neuroticism were higher in the HRS, whereas the other factors (i.e., maternal education, history of renting, positive interactions with friends and hopelessness) were higher in MIDUS, although the direction of association only differed meaningfully for neuroticism. Methods for the replication are included in the *SI Appendix*.

Discussion

In the present study, factors from across the behavioral and social sciences were explored as independent predictors of mortality, and many were discovered as important regardless of the applied statistical approach. These included behaviors (e.g., smoking, alcohol abuse, physical inactivity), financial wellbeing (e.g., reported financial difficulties or lower wealth), social experiences (e.g., divorce, negative interactions with children, discrimination), and psychological characteristics (e.g., trait negative affectivity, lower life satisfaction). In addition, we found that each overall domain contributed substantially to explaining the variance in time to death. Because of different strengths of the complementary analytic approaches, some predictors were better identified with a particular approach. While adverse psychosocial experiences in childhood were associated with mortality in independent Cox regression analysis, its importance ranked higher in the lasso model. Conversely, differences were apparent for discrimination, lower life satisfaction, and sleep problems, and, although important when examined independently, they reduced in strength when considered in a model simultaneously with recent financial difficulties, health behaviors, negative social interactions with children, and early psychosocial adversity. These findings provide evidence to support the suggestion that other prospective and national cohort studies should widen the net that is cast when testing behavioral and social factors by including these and related measures of social experiences and psychological characteristics from across the life course.

There were some individual variables that were not predictive of mortality across domains in the individual Cox regression analyses and machine learning algorithms. For example, childhood economic difficulties and religiosity were unrelated with any approach

There were some findings that ran counter to those in the literature. In the lasso regression analysis, older adults whose fathers were unemployed when they were children or whose mothers had lower education levels were more likely to live longer. These results may result from a cohort effect for adults born in the 1930s, during the Great Depression, and whose fathers were sent to Europe to fight in World War II. As adults in the HRS continue to age, we can potentially examine the effects of childhood socioeconomic conditions as a function of the different generations of children included in the study. Another finding that seems counterintuitive in the lasso regression is that increased adversity in adulthood was related to reduced risk of mortality. Some research suggests that moderate levels of psychosocial adversity may predict lower distress across the lifespan compared to no adversity or high adversity (25). In the current investigation, we did not develop our aims to examine the curvilinear relationship between our selected factors, including adulthood adversity, with mortality. Our current findings provide fruitful grounds for future explorations. These counterintuitive results may also be due to mortality selection.

There are several factors not included in the current study that were either not measured in the HRS or purposefully excluded from our analyses. The HRS does not have data on dietary consumption, nor measures of environmental contaminants, which have been the focus of prior prediction models of mortality (4, 18). We excluded genetic and health markers because our primary focus was to examine predictors typically examined in the behavioral and social sciences (12). A recent investigation by Liu and colleagues (26) using data from the HRS determined that genetic factors, health behaviors, and life course circumstances accounted for one third of the variation of a novel summary marker of health calculated from nine biomarkers. Importantly, there was also a significant gene–environment interaction by which cardiovascular disease polygenic risk scores and socioeconomic disadvantage compounded accelerated aging. While the study did not use similar measures to ours—we included social connection with family and friends and personality characteristics—Liu's study demonstrates the added information that can be gleaned when including genetic markers in studies of health and mortality. Liu and colleagues' study also highlights that each investigated domain accounts for significant variability in the selected outcome, similar to our investigation.

Future studies should develop lifespan models that include genetic, biological, and health markers in addition to the individual-level factors considered here (e.g., economic, behavioral, social, and psychological) and more macrolevel markers (e.g., built, natural, and social environments, including structural racism, neighborhood cohesion, policies at state/provincial and national levels) that have previously been demonstrated as determinants of health and mortality. Social trajectory theories (27–31) highlight the pathways through which social and economic experiences early in life and through adulthood foster health habit formation, economic attainment, and psychological characteristics that ultimately shape the healthspan and life expectancy. While our investigation did not demonstrate that early economic experiences impact mortality later in life, social trajectory research is clear on the importance of early economic experiences in laying the ground for economic attainment, health habits, and personality in midlife and later in life. Importantly, those adults who emerged from poor economic conditions earlier in life may have died at earlier ages and, resultingly, may not have been included in the HRS or our analyses with the HRS data (i.e., selective mortality). Lifespan modeling using longitudinal prospective studies that started collecting data with younger participants than the HRS and that includes genetic, built, and political environment measures would allow for testing the role of early social and economic conditions in laying the foundation for social and economic attainment in adulthood, health habits, and personality formation.

Our study has several limitations. It is important to note that our study design does not allow for causal interpretations. Although we can rank which factors best predict mortality from across disciplines, we cannot assert that modifying levels of these factors would change an individual's mortality risk. The utility of our approach is the ranking of multiple factors from across the disciplines, which can expand future considerations of what types of predictors should be more thoroughly investigated, and is intended as a means of hypothesis generation for future observational and clinical studies. Second, our measures of adulthood psychosocial adversities were limited in scope and number, and do not include the full breadth and number of events that are

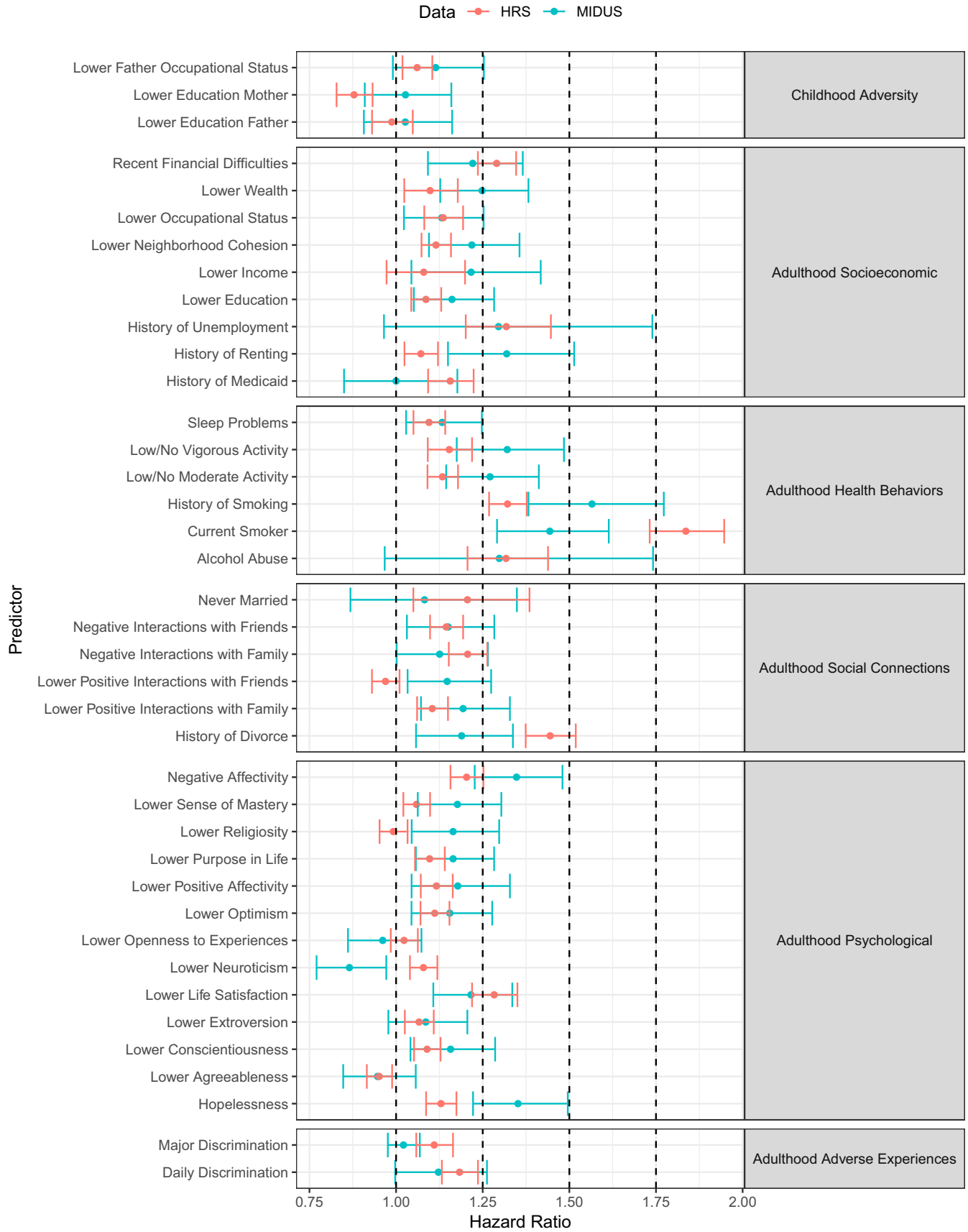

Fig. 3. Independent Cox regression hazard ratios and confidence intervals (unweighted) of each predictor for mortality in the HRS (red) and MIDUS (blue) studies with 39 harmonized variables from the 6 domains. Confidence intervals that include 1 indicate that a predictor is not statistically significant at the 5% level. Results are not corrected for multiple tests. Age is used as the baseline hazard. Larger hazard ratios indicate higher mortality risk.

often measured by life event scales (32). As noted elsewhere, other relevant adversities in adulthood (e.g., food insecurity, domestic abuse) were not included in the HRS (33). Third, there may be some concern that the reporting of childhood events is prone to reporting and recall error (34, 35), as seen in previous studies where early abuse experiences are underreported (36, 37). Hardt and colleagues (37, 38) have noted, however, that, when adverse experiences are clearly defined, as in the case in our report, reporting bias should have little effect in studies. Nationally representative studies have supported Hardt's conclusion, demonstrating that rates and health impacts of childhood health, economic, and traumatic experiences reported as adults are similar to prospective cohort studies that commenced in childhood (39–41). Our results are specific to 6 y of follow-up time for mortality, and the factors we identify could differ with longer follow-up time.

We replicated our findings in the national Midlife in the United States Study, which was selected for its breadth of psychological factors, its location (i.e., the United States), the capacity to compare similar age groups, and access to linked mortality data. There may also be utility in comparing how factors across the behavioral and social sciences predict mortality compared to morbidity or physical functioning assessed with objective data collected in electronic health records. Replication should be further considered with nationally representative datasets from across the globe, including the UK Biobank and the International Sister Studies, of which the HRS is just one. These nationally representative sister studies have been conducted in Brazil, China, Costa Rica, England, Europe, India, Indonesia, Ireland, Japan, Korea, Malaysia, Mexico, Northern Ireland, South Africa, Scotland, and Thailand, and collected data from these studies have been harmonized through the effort to create the Gateway to Global Aging Data platform. Establishing global and country-specific behavioral and social predictors of mortality across these developed and developing countries may expand our understanding of the challenges faced globally by an aging population as well as clarify unique challenges facing specific countries.

The purpose of our investigation was to draw attention to specific early life and adulthood socioeconomic, behavioral, social, and psychological factors that may impact mortality in a nationally representative and prospective study in the United States. Our goal was to expand the types of factors that current and future investigations could consider in their methodologies in order to obtain a more comprehensive understanding of the impact of behavioral and social factors on health and mortality. Similar investigations with other national studies may help consolidate around a critical set of behavioral, economic, social, and psychological factors to include in future studies.

Methods

Predictors. Fifty-seven predictors measured from years 1992 to 2008 across the following domains were measured: (*i*) adverse socioeconomic and psychosocial experiences during childhood, (*ii*) socioeconomic conditions, (*iii*) health behaviors, (*iii*) social connections, (*v*) psychological characteristics, and (*vi*) adverse experiences during adulthood. A full list of the 57 variables in the current analysis is provided in Table 1. *SI Appendix, Methods and Materials*, includes further details of the measures used and how data were processed for the current analysis.

Outcome. Mortality records in the HRS were linked to National Death Index (NDI) through 2011. Beyond 2011, mortality information was ascertained through exit interviews obtained from a family member of the respondent. More information is provided in ref. 42.

Data Source. All data are publicly available and can be retrieved upon request from https://hrs.isr.umich.edu/data-products/access-to-public-data. Analyses codes are included in Datasets S1 and S2.

Statistical Analysis

Means and SDs were calculated for each continuous variable, and totals and percent for each categorical variable. Correlations among all risk factors are presented in *SI Appendix*, Fig. S2. The analyses were weighted to account for the survey design of the Health and Retirement Study. The outcome of interest across our analyses was time to death. Multivariate Cox regression analyses (43) were used to examine the contribution of each predictor to the hazard of death, with age as the time unit. Continuous variables were standardized. Binary variables were coded −1 and 1, and categorical variables with three categories, −1, 0, and 1. These were implemented using the "survival" package in R (44). For Fig. 1, predictors were examined independently in multivariate Cox models adjusted for gender, race, ethnicity, and migrant status. We estimated cluster-robust SEs by household due to the sampling of respondents and their spouses in the HRS. We obtained standardized hazard ratios (HRs) and 95% confidence intervals for each predictor, which were corrected for multiple comparisons using Bonferroni correction for 57 comparisons (45). We also completed follow-up sensitivity multivariate Cox regression analyses excluding those who died within 2 y from 2008 and with an interaction between each predictor and gender [male versus female, racial identification (white versus nonwhite), educational status (completed high school or not), and age (<75 y versus 75+ y)].

Next, to estimate the proportion of variance explained by each predictor domain, we estimated a multivariate Cox regression model for each domain independently, adjusted for sex, race, ethnicity, and migrant status, and obtained the pseudo–R-squared for each domain. We used exploratory principal components analyses to reduce the dimensionality of our data. This process was completed in two steps. First, the imputePCA function from the missMDA package was used to impute any missing values. Then, for each category of variable, the PCA function from FactoMineR was used to identify the dimensions within that category. The three dimensions which explained the greatest amount of variance were used in the multivariate Cox regression models for each domain independently and the domains combined.

Next, we performed lasso regression (22) in order to fit a prediction model which selects the best fitting parsimonious model given a large number of potential predictors of mortality. Given the time-to-event nature of the data, we fit our models using the glmnet() function in R with family="cox". Lasso is a generalized linear model that is fit with penalized maximum likelihood and is appropriate when correlations are not high enough to cause any issues with multicollinearity, as is the case in the current study. It was developed to provide a variable selection approach which has the advantage, as compared to backward or forward selection models, of penalizing coefficients rather than completely dropping or adding variables to the model, even though variables can be set to zero if they do not add sufficiently to the model (46). We used cross-validation with the cv.glmnet function to select a model with the minimum mean cross-validation error.

Finally, we used random forest survival analysis to predict survival over the survey period, using the domain-specific predictors and age as the baseline hazard. The random forest algorithm is a nonparametric, ensemble machine learning tool first introduced by Breiman (47) as an extension of classification and regression trees (CART) and bagging.

The random forest algorithm works by repeatedly drawing bootstrap samples from the original sample and a random selection of predictors to grow a predetermined number of

decision trees across which results are pooled. A training data set consisting of n of N cases (two thirds of the original sample) is generated for each of k decision trees, and the remaining cases (one third of the original sample) are used as test data to estimate the out of bag (OOB) classification error. A random sample m of M predictors is selected at each node, and the one predictor that best discriminates discrepancies in survival is chosen for that particular split. As a result, the root node of each decision tree represents the strongest predictor, and the splits that follow are based on the successively strongest predictors. A final classification is made using a majority of votes across all trees.

We implemented random forest survival analysis with all predictors included in a single model. Classification was performed using the "rfsrc" function from the R package "randomForestSRC v2.8.0" (48, 49) with a maximum of 5 iterations, each with 500 classification trees. We obtained variable importance plots for each predictor. Missing values were imputed using missForest (50).

We evaluated the extent to which each variable contributes to predicting survival using the metric of variable importance using the vimp function from the R package randomForestSRC v2.8.0 (48, 49). We calculate variable importance using random permutation of the variable approach. This is done by comparing the accuracy of prediction of each tree as estimated from the model with the tree with each particular variable permuted. A summary of these differences across all of the trees is used to calculate variable importance (51). The figure shows positive variable importance in blue, which indicates this factor increases the predictive nature of the model. Negative variable importance is shown in red, indicating these variables decrease the accuracy of prediction. This is likely due to random error. Note that the variable importance scale is quite low, meaning that, based on the random forest model with baseline hazard of age, these variables are not highly predictive of survival relative to age in this model.

ACKNOWLEDGMENTS. The HRS is supported by the National Institute on Aging (NIA; U01 AG009740) and the Social Security Administration. The Midlife in the United States (MIDUS) investigation was supported by NIA Grants P01-AG020166 and R01-AG019239. The original study was supported by the John D. and Catherine T. MacArthur Foundation Research Network on Successful Midlife Development. The funding sources had no involvement in the study design; data collection, analysis, or interpretation; nor the writing and submission of this article. This research was supported by the Canada Research Chairs program (E.P.) and a Population Research Training Grant (NIH T32 HD007242) awarded to the Population Studies Center at the University of Pennsylvania by the National Institutes of Health's (NIH)'s Eunice Kennedy Shriver National Institute of Child Health and Human Development (J.W.). D.H.R. was supported by the National Institute on Aging at the National Institutes of Health (K01AG047280). This work was supported by the National Institute on Aging at the National Institutes of Health with a grant to the Stress Measurement Network (R24AG048024).

1. M. Mather, L. Jacobsen, K. Pollard, Aging in the United States. *Popul. Bull.* **70** (2015).
2. P. L. Mabry, D. H. Olster, G. D. Morgan, D. B. Abrams, Interdisciplinarity and systems science to improve population health: A view from the NIH Office of behavioral and social sciences research. *Am. J. Prev. Med.* **35** (suppl. 2), S211–S224 (2008).
3. V. Kontis *et al*., Future life expectancy in 35 industrialised countries: Projections with a Bayesian model ensemble. *Lancet* **389**, 1323–1335 (2017).
4. J. M. McGinnis, W. H. Foege, Actual causes of death in the United States. *JAMA* **270**, 2207–2212 (1993).
5. A. H. Mokdad, J. S. Marks, D. F. Stroup, J. L. Gerberding, Actual causes of death in the United States, 2000. *JAMA* **291**, 1238–1245 (2004).
6. D. W. Brown *et al*., Adverse childhood experiences and the risk of premature mortality. *Am. J. Prev. Med.* **37**, 389–396 (2009).
7. A. C. Carlsson *et al*., Financial stress in late adulthood and diverse risks of incident cardiovascular disease and all-cause mortality in women and men. *BMC Public Health* **14**, 17 (2014).
8. J. Holt-Lunstad, T. B. Smith, J. B. Layton, Social relationships and mortality risk: A meta-analytic review. *PLoS Med.* **7**, e1000316 (2010).
9. M. Jokela *et al*., Personality and all-cause mortality: Individual-participant meta-analysis of 3,947 deaths in 76,150 adults. *Am. J. Epidemiol.* **178**, 667–675 (2013).
10. C. R. Gale *et al*., When is higher neuroticism protective against death? Findings from UK biobank. *Psychol. Sci.* **28**, 1345–1357 (2017).
11. L. L. Barnes *et al*., Perceived discrimination and mortality in a population-based study of older adults. *Am. J. Public Health* **98**, 1241–1247 (2008).
12. D. Fanelli, J. P. A. Ioannidis, US studies may overestimate effect sizes in softer research. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15031–15036 (2013).
13. A. Ganna, E. Ingelsson, 5 year mortality predictors in 498,103 UK biobank participants: A prospective population-based study. *Lancet* **386**, 533–540 (2015).
14. C. J. Patel, J. P. A. Ioannidis, Studying the elusive environment in large scale. *JAMA* **311**, 2173–2174 (2014).
15. L. C. Pilling *et al*., Human longevity: 25 genetic loci associated in 389,166 UK biobank participants. *Aging (Albany NY)* **9**, 2504–2520 (2017).
16. A. Ganna *et al*., Genetic determinants of mortality. Can findings from genome-wide association studies explain variation in human mortality? *Hum. Genet.* **132**, 553–561 (2013).
17. C. J. Patel, J. Bhattacharya, A. J. Butte, An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* **5**, e10746 (2010).
18. C. J. Patel *et al*., Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the United States national health and nutrition examination survey. *Int. J. Epidemiol.* **42**, 1795–1810 (2013).
19. X. Zhuang *et al*., Environment-wide association study to identify novel factors associated with peripheral arterial disease: Evidence from the National Health and Nutrition Examination Survey (1999-2004). *Atherosclerosis* **269**, 172–177 (2018).
20. G. M. Buck Louis, R. Sundaram, Exposome: Time for transformative research. *Stat. Med.* **31**, 2569–2575 (2012).
21. M. M. Glymour, T. L. Osypuk, D. H. Rehkopf, Invited commentary: Off-roading with social epidemiology—exploration, causation, translation. *Am. J. Epidemiol.* **178**, 858–863 (2013).
22. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
23. S. Aichele, P. Rabbitt, P. Ghisletta, Think fast, feel fine, live long: A 29-year study of cognition, health, and survival in middle-aged and older adults. *Psychol. Sci.* **27**, 518–529 (2016).
24. J. P. Shaffer, Multiple hypothesis testing. *Annu. Rev. Psychol.* **46**, 561–584 (1995).
25. M. D. Seery, E. A. Holman, R. C. Silver, Whatever does not kill us: Cumulative lifetime adversity, vulnerability, and resilience. *J. Pers. Soc. Psychol.* **99**, 1025–1041 (2010).
26. Z. Liu *et al*., Associations of genetics, behaviors, and life course circumstances with a novel aging and healthspan measure: Evidence from the Health and Retirement Study. *PLoS Med.* **16**, e1002827 (2019).
27. D. Umberson, R. Crosnoe, C. Reczek, Social relationships and health behavior across life course. *Annu. Rev. Sociol.* **36**, 139–157 (2010).
28. S. Liu, R. N. Jones, M. M. Glymour, Implications of lifecourse epidemiology for research on determinants of adult disease. *Public Health Rev.* **32**, 489–511 (2010).
29. L. F. Berkman, Social epidemiology: Social determinants of health in the United States: Are we losing ground? *Annu. Rev. Public Health* **30**, 27–41 (2009).
30. S. Oishi, J. Graham, Social ecology. *Perspect. Psychol. Sci.* **5**, 356–377 (2010).
31. O. Solar, A. Irwin; World Health Organization, A conceptual framework for action on the social determinants of health paper 2 (Policy and Practice). (WHO, Geneva, Switzerland, 2010).
32. G. Brown, T. Harris, "Life events and measurement" in *Life Events and Illness*, G. Brown, T. Harris, Eds. (The Guilford Press, 1989), pp. 3–45.
33. E. Puterman *et al*., Lifespan adversity and later adulthood telomere length in the nationally representative US Health and Retirement Study. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E6335–E6342 (2016).
34. B. Maughan, M. Rutter, Retrospective reporting of childhood adversity: Issues in assessing long-term recall. *J. Pers. Disord.* **11**, 19–33 (1997).
35. S. R. Dube, D. F. Williamson, T. Thompson, V. J. Felitti, R. F. Anda, Assessing the reliability of retrospective reports of adverse childhood experiences among adult HMO members attending a primary care clinic. *Child Abuse Negl.* **28**, 729–737 (2004).
36. K. MacDonald *et al*., Minimization of childhood maltreatment is common and consequential: Results from a large, multinational sample using the childhood trauma questionnaire. *PLoS One* **11**, e0146058 (2016).
37. J. Hardt, M. Rutter, Validity of adult retrospective reports of adverse childhood experiences: Review of the evidence. *J. Child Psychol. Psychiatry* **45**, 260–273 (2004).
38. J. Hardt, A. Sidor, M. Bracko, U. T. Egle, Reliability of retrospective assessments of childhood experiences in Germany. *J. Nerv. Ment. Dis.* **194**, 676–683 (2006).
39. J. P. Smith, Reconstructing childhood health histories. *Demography* **46**, 387–403 (2009).
40. E. Havari, F. Mazzonna, Can we trust older people's statements on their childhood circumstances? Evidence from SHARELIFE. *Eur. J. Popul.* **31**, 233–257 (2015).

41. J. Hardt, P. Vellaisamy, I. Schoon, Sequelae of prospective versus retrospective reports of adverse childhood experiences. *Psychol. Rep.* **107**, 425–440 (2010).
42. D. Weir, Validating mortality ascertainment in the health and retirement study. https://hrs.isr.umich.edu/publications/biblio/9022. Accessed 29 August 2019.
43. D. R. Cox, *Regression Models and Life-Tables*, (Springer, New York, NY, 1992), pp. 527–541.
44. T. M. Therneau, T. Lumley, Analysis P. on CRAN, Undefined 2014, Package "Survival." Version 3.1.8. https://cran.r-project.org/web/packages/survival/survival.pdf. Accessed 5 August 2019.
45. R. J. Simes, An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754 (1986).
46. T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations. https://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS_corrected_1.4.16.pdf. Accessed 5 August 2019.
47. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
48. C. Strobl, T. Hothorn, A. Zeileis, Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package (2009). https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf. Accessed 24 April 2017.
49. H. Ishwaran, U. Kogalur, M. Kogalur, "Package 'randomForestSRC,'" Version 2.8.0. https://cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf. Accessed 5 June 2020.
50. D. J. Stekhoven, P. Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
51. H. Ishwaran, Variable importance in binary regression trees and forests. *Electron. J. Stat.* **1**, 519–537 (2007).