RESEARCH ARTICLE

Structured Inequality, Uncertain Lifespans: Demographic Perspectives on Predicting Individual-Level Longevity

CASEY F. BREEN AND NATHAN SELTZER

*There are striking disparities in life expectancy across sociodemographic groups in the United States, shaped by structural forces such as racism, class inequality, and policy environments. To what extent do sociodemographic characteristics structure—or fail to structure—individual lifespans? Using U.S. Census data linked to administrative death records, we assess how well early-adulthood social, economic, and demographic characteristics predict individual lifespan in a cohort of men born in 1910 and observed through their deaths between 1975 and 2005 (*N = *121,000). Despite large group-level disparities, we find that sociodemographic characteristics measured in early adulthood explain less than two percent of the overall variation in individual lifespan. These findings reaffirm a central demographic regularity: variance in life expectancy between groups is small compared to variation in lifespan within groups. This highlights the fundamentally nondeterministic nature of how structural inequality shapes individual mortality.*

Introduction

Population researchers have made substantial progress in understanding the contours, disparities, and determinants of mortality in the United States (Elo 2009; Gutin and Hummer 2021; Dowd, Polizzi, and Tilstra 2025). An extensive body of work has applied classic demographic methods to analyze levels, trends, and inequalities in mortality along racial and socioeconomic lines (Wrigley-Field 2025; Schwandt et al. 2021). In parallel, important advances have been made in theorizing the social origins of

Casey F. Breen, Department of Sociology, Population Research Center, and Center for Aging and Population Sciences, University of Texas at Austin, Austin, TX, 78712, USA. E-mail: casey.breen@austin.utexas.edu. Nathan Seltzer, Department of Demography, University of California, Berkeley, Berkeley, CA, 94720, USA.

mortality disparities (Hayward and Gorman 2004; van Raalte 2021; Link and Phelan 1995; Dannefer 2003) and identifying causal determinants of longevity (Fletcher and Noghanibehambari 2024; Cutler, Deaton, and Lleras-Muney 2006; Chetty and Hendren 2018).

Against the backdrop of descriptive, causal, and theoretical work on mortality, the explosion of rich microdata has enabled a new *predictive perspective*. In this study, we apply this perspective to ask: Can observable sociodemographic characteristics—such as education, income, race, and marital status—predict individual-level lifespan? This question tests the limits of social determinism. If lifespan is highly predictable, it would suggest that one's lifespan is tightly structured by systemic forces ("demography is destiny"). Low predictability would imply that, despite large and well-documented between-group mortality disparities, most lifespan variation remains unexplained by major social or economic factors. This distinction matters for how we think about lifespan inequality: Is mortality governed more by structural inequalities or by stochastic individual variation?

To evaluate this, we analyze linked U.S. Census and Social Security mortality records for over 121,000 men born in 1910, following a single cohort from early adulthood through death to isolate variation in lifespan among individuals observed from a common baseline age. We observe large between-group disparities for this cohort, with gaps in life expectancy at age 65 of nearly three years between those with high and low education. Yet our predictive models only explain 1.3 percent of the variance in lifespan. This highlights a core tension: there are substantial and important between-group disparities, but individual-level variation is great enough that we cannot use these sociodemographic characteristics to accurately predict lifespan for a given person. In other words, predictability is low not because between-group inequality is absent, but because within-group individual variation dominates individual lifespan (Vaupel 1988; Caswell 2023; van Raalte et al. 2012).

Mortality demography, at its core, is concerned with population rates and group-level disparities. What does a predictive perspective add? First, using predictive performance as a diagnostic tool quantifies how much of the variation in lifespan within a cohort is captured by sociodemographic indicators of structural advantage and disadvantage. This approach generalizes traditional decompositions of lifespan variation, which in most empirical applications have examined one categorical covariate at a time, by translating multiple covariates into measures of explanatory power at the individual level. In short, prediction reframes classical decomposition as an inquiry into individual-level explanatory power, leveraging both categorical and continuous characteristics. Second, limited predictability reveals how inequality and uncertainty coexist: stochasticity shapes the overall distribution of lifespans, while structural disparities shift its mean and skewness but remain limited in their ability to determine individual outcomes. Finally,

the low predictability of longevity based on sociodemographic characteristics is itself demographically meaningful. It provides empirical insight into a universal form of uncertainty—one that individuals consider when making life decisions (e.g., Will I outlive my retirement savings?). This perspective aligns with the emerging field of uncertainty demography (Trinitapoli 2023), which calls for making uncertainty more central to demographic inquiry and highlights the distinctive capacity of demographic approaches in illuminating the role of uncertainty in social life.[1]

Background

Social scientists studying mortality have a long-standing interest in describing aggregate disparities in mortality. There are striking class-based (Elo 2009; Montez, Hummer, and Hayward 2012; Chetty et al. 2016), racial (Hummer, Benjamins, and Rogers 2004; Hayward and Heron 1999; Feigenbaum, Muller, and Wrigley-Field 2019; Wrigley-Field 2020), and geographic (Dowd et al. 2024; Montez, Harward, and Wolf 2017) disparities in mortality. While overall longevity has increased over the course of the 20th century (Dowd, Polizzi, and Tilstra 2025), inequality in mortality has also increased over time in the United States along key dimensions (Preston and Elo 1995). Researchers have also documented paradoxical or surprising mortality dynamics, including mortality crossovers (Vaupel, Manton, and Stallard 1979; Wrigley-Field 2020) and the Hispanic mortality paradox (Hummer 2000; Elo et al. 2004; Lariscy, Hummer, and Hayward 2015).

A particularly relevant line of research decomposes variance in longevity into *between-group* and *within-group* contributions (van Raalte et al. 2012; Caswell 2023; Steiner, Tuljapurkar, and Orzack 2010). Between-group variation reflects systematic differences in life expectancy across groups defined by characteristics such as race, class, or geography. This is sometimes referred to as heterogeneity in the mortality literature (Caswell 2023). In contrast, within-group variation captures the differences in lifespan among individuals who share similar risk profiles. This is sometimes referred to as individual stochasticity (Caswell 2009), dynamic heterogeneity (Snyder and Ellner 2018), intragroup heterogeneity (Permanyer, Sasson, and Villavicencio 2023), and luck (Steiner, Tuljapurkar, and Orzack 2010). This is where chance processes, not observed characteristics, drive variation in lifespan among individuals with identical mortality risk profiles.[2] Both between-group and within-group variation jointly contribute to overall lifespan variation.

This conceptual distinction has motivated recent empirical work quantifying the relative contribution of between-group and within-group variation to overall lifespan variation (Seaman, Riffe, and Caswell 2019; Permanyer et al. 2018). These studies suggest that within-group variation dominates, with characteristics such as income, education, or neighborhood

deprivation only explaining a small fraction of overall lifespan variation (Caswell 2023). Most applications of variance decomposition in demography have focused on single categorical factors considered separately; however, recent methodological extensions allow multiple covariates and their interactions to be incorporated within decomposition frameworks (Caswell and Van Daalen 2025). Our approach complements this line of work by shifting from variance partitioning toward individual-level prediction using high-dimensional covariate sets.

Prior research on mortality prediction

Improvements in data and computation have made individual-level prediction a more feasible research goal (Kashyap 2021), and prediction is increasingly seen as a valuable tool for social science research (Hofman et al. 2021). Several studies have applied a *predictive perspective* to estimate individual mortality (Einav et al. 2018; Rose 2013; Badolato et al. 2026; Savcisens et al. 2024) and to identify the most important predictors of survival (Goldman, Glei, and Weinstein 2016;2017; Puterman et al. 2020).[3] These efforts align with broader methodological shifts in the social sciences and efforts to predict life outcomes (Zheng and Cheng 2025; Salganik et al. 2020).

Studies of individual mortality prediction can be broadly classified as follows: (1) studies that predict all-cause mortality using sociodemographic, behavioral, and health variables from surveys or administrative data; and (2) clinical studies that predict short-term mortality based on diagnoses, symptoms, and other patient-level information.[4] Here, we focus on studies predicting all-cause mortality (Table 1).

To date, mortality prediction studies have used a period-based design. In this design, researchers pool individuals of different ages and predict their survival over a fixed horizon, addressing questions such as, "Will this individual die in the next five years?" For example, Rose (2013) used data from an aging cohort in Sonoma County to predict five-year mortality from physical activity, smoking, self-rated health, and age, achieving moderate predictive power ($R^2 = 0.201$). Badolato et al. (2026) used data from the Health and Retirement Study to estimate mortality hazard functions, pooling person-wave observations across survey years. They find overall predictive accuracy to be low, especially for men, non-Hispanic Blacks, and individuals with low education. Across all models, age is by far the strongest predictor.

Even with detailed medical data, mortality prediction efforts report mixed success. Einav et al. (2018) used Medicare claims to predict 12-month mortality for elderly patients and found that even the highest risk individuals had less than a 25 percent chance of dying, suggesting that mortality is difficult to predict even with detailed health information. In a prominent empirical example, Savcisens et al. (2024) used Danish registry data and a

TABLE 1 Summary of past studies on individual-level prediction of all-cause mortality

Study	Data source	Covariates	Sample size
Puterman et al. (2020)	Health and Retirement Study	Sociodemographic, behavioral, and health	39,248
Badolato et al. (2026)	Health and Retirement Study	Sociodemographic, behavioral, and health	39,248
Savcisens et al. (2024)	Danish Registry Data	Sociodemographic, behavioral, and health	100,000
Rose (2013)	Study of Physical Performance and Age-Related Changes in Sonoma	Sociodemographic, behavioral, and health	2,092
Einav et al. (2018)	Medicare enrollees	Sociodemographic, behavioral, and electronic medical records	5,631,168
This study	CenSoc (linked census and Social Security Mortality records)	Sociodemographic	121,000

large language model to predict premature mortality among adults aged 30–60 over a four-year window. The authors reported mixed success even using age as a covariate, with their best performing model achieving a corrected Matthews correlation coefficient of 0.41, indicating only modest ability to classify deaths.

Period-based approaches are well-suited to many practical forecasting exercises that assess mortality risk across a population heterogeneous in age. Although even the most comprehensive studies have found clear limits to predicting lifespan (Badolato et al. 2026), such predictive mortality models have plausible use cases: for example, identifying individuals at the highest short-term risk of death.[5] In a period-based design, models generally rely on age as the key predictor. Because mortality risk rises steeply with age, models can achieve moderate predictive performance simply by inferring that older individuals are more likely to die. Even when age is excluded, many covariates (e.g., self-rated health or homeownership status) proxy for age, meaning that much of predictive performance hinges on chronological aging rather than social or structural factors.

In contrast, a cohort perspective offers a cleaner test of how much variation in lifespan measurable characteristics can explain. This perspective asks: "At what age will a given member of a birth cohort die?" By following individuals born in the same year, this design holds age constant and absorbs cohort-level effects, isolating the contribution of sociodemographic factors to within-cohort variation in lifespan. To our knowledge, no prior

study has applied a cohort-based prediction framework to quantify lifespan predictability in the United States.[6] Applying this perspective, we quantify how sociodemographic characteristics structure within-cohort variation in individual longevity.

Data and study design

In this study, we use the publicly available CenSoc-DMF dataset (Goldstein et al. 2021). This dataset links the full-count U.S. 1940 Census (Ruggles et al. 2020) with mortality records from the Social Security Death Master File (DMF). The DMF captures nearly all deaths in the United States over age 55 between 1975 and 2005 (Alexander 2018; Hill 2001). The 1940 Census provides individual-level sociodemographic characteristics that we use as predictors, including educational attainment, race, income, housing tenure, occupation, and marital status. The CenSoc-DMF file does not include women, as surname changes during marriage preclude reliable record linkage between the 1940 Census and Social Security mortality records. For more technical details on data linkage and methodology, see Breen, Osborne, and Goldstein (2023).

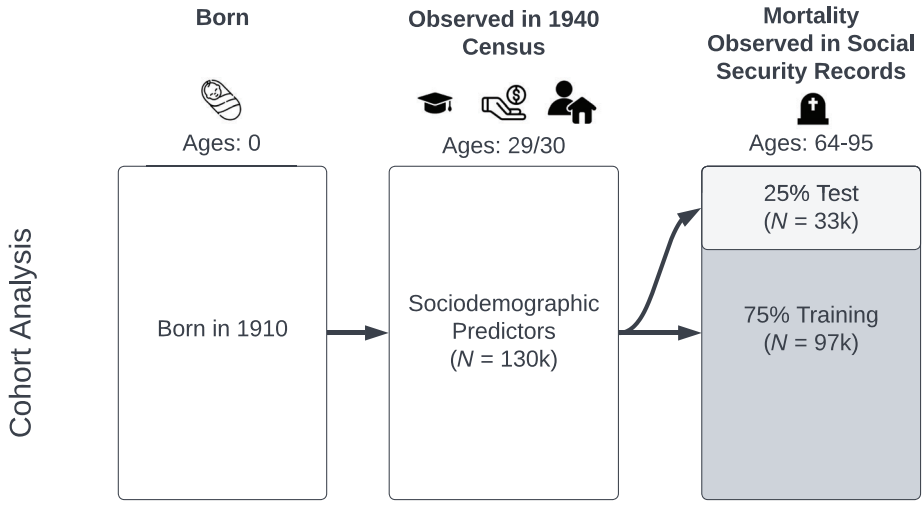
The CenSoc-DMF dataset offers three key advantages for mortality prediction. First, it enables the study of real birth cohorts tracked longitudinally over a 30-year mortality observation window. Second, its large size provides sufficient statistical power to analyze individual birth cohorts. Finally, it includes the core sociodemographic characteristics that are standard indicators of structural inequality.

For our analysis, we focus on individuals born in 1910, who were observed in the 1940 Census at age 29 or 30 and died in our mortality observation window between ages 64 and 95. This sample is illustrated in Figure 1. By restricting the analysis to a single birth cohort, we hold age constant and isolate variation in lifespan. We focus on this cohort, who by 1940 had largely finished their education and entered the labor force, offering a stable snapshot of early-adult socioeconomic conditions.

To assess representativeness, we compare the composition of our linked 1910 birth cohort sample to the composition of all men in the 1940 Census born in 1910. As shown in Figure 2, our linked sample is broadly representative of the corresponding birth cohort observed in the 1940 Census. Like most historical linked samples, it slightly overrepresents individuals of higher socioeconomic status and White individuals. Because our focus is on assessing predictive accuracy, not making population-level inferences, this slight compositional difference is unlikely to affect our results.

All 1940 Census covariates plausibly related to mortality were included as predictors (see Table S1 in the Supporting Information). Categorical covariates (e.g., race, marital status) were dichotomized, and continuous covariates were standardized to a common scale, with a mean of 0 and

FIGURE 1 Overview of our analytic sample. We observe their early-adulthood characteristics in the 1940 Census at age 29 or 30, and their mortality between ages 64 and 95



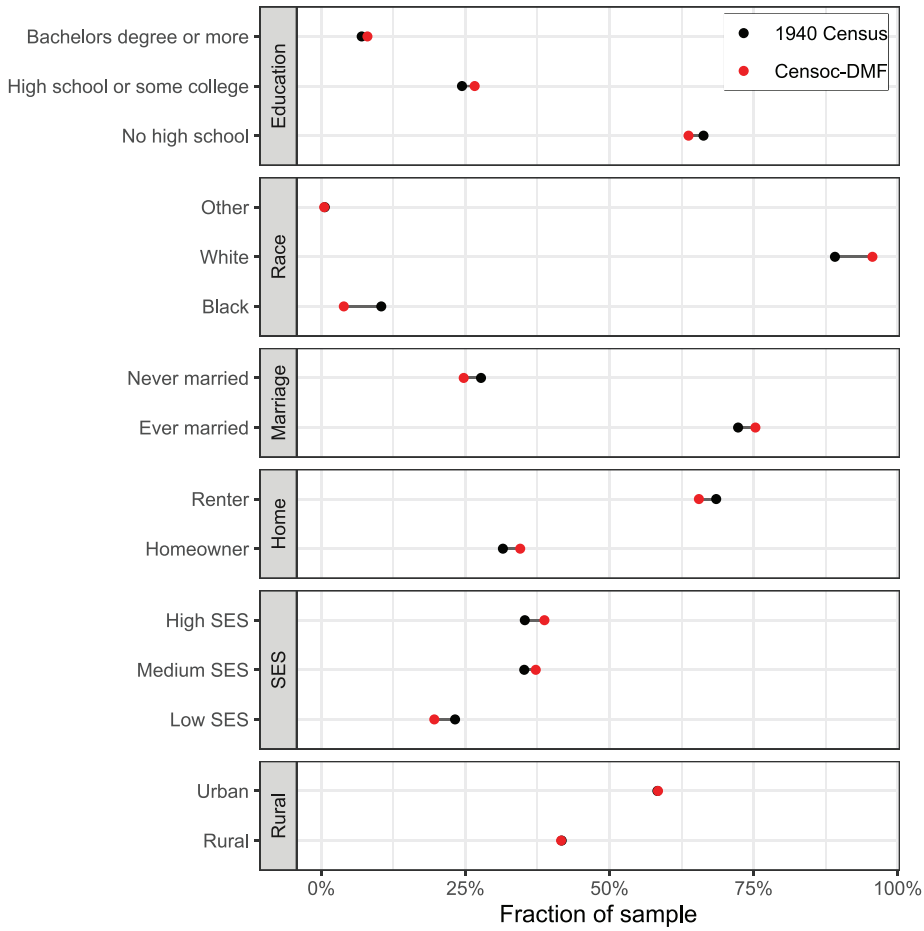
a standard deviation of 1. The outcome measure is age at death in years, and we exclude cases with missing data.[7] These predictors are measured cross-sectionally at ages 29 or 30 and do not capture subsequent changes over the life course.

Methods

We use a machine learning approach to assess how well an individual's lifespan can be predicted from sociodemographic characteristics. Flexible algorithms are well-suited to this task because they can detect nonlinearities and interactions among correlated predictors (Lundberg, Brand, and Jeon 2022). We implement an ensemble Superlearner (Van der Laan, Polley, and Hubbard 2007), which combines multiple algorithms to improve predictive accuracy and reduce overfitting. Specifically, the Superlearner combines predictions from multiple algorithms, weighting them according to their performance.[8]

We randomly split the sample into a training partition (75 percent) and a holdout partition (25 percent).[9] The training set includes all predictors along with our outcome of interest, age at death in years. We use this training data to fit the full set of machine learning algorithms. To evaluate predictive performance, we apply the trained algorithms to the holdout set, withholding information on actual ages at death. We then compare the predicted ages at death with the true, withheld outcomes to assess model accuracy. Additional implementation details, including algorithm specifications and validation procedures, are provided in Appendix Section B of the

FIGURE 2 Each facet compares the composition of our analytic sample (red) with that of the 1940 U.S. Census (black) for a given covariate. Overall, the sample composition aligns closely with the population

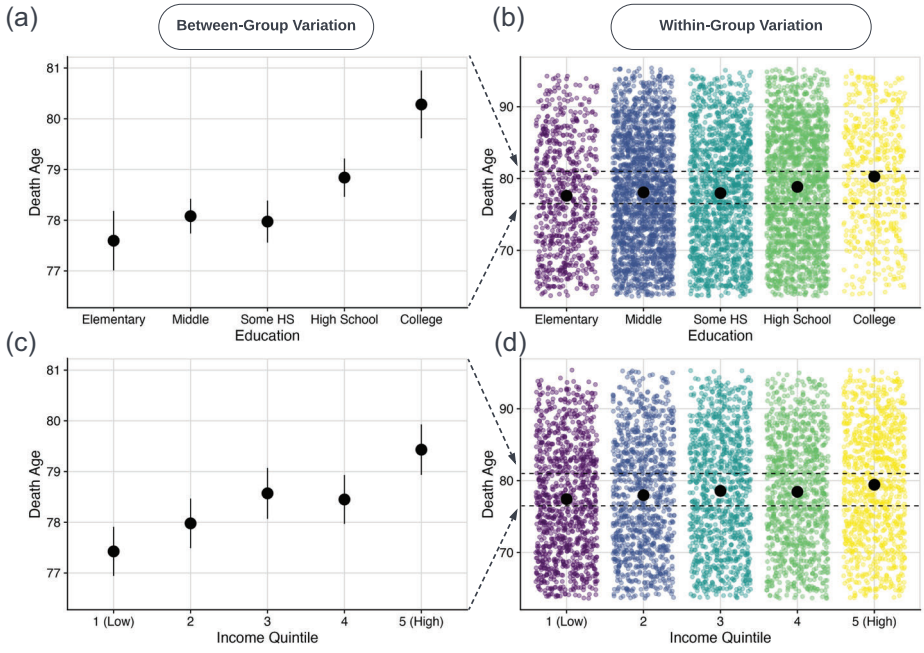


Supporting Information. Our completed REFORM checklist, a reproducibility and transparency framework for machine learning (Kapoor et al. 2024), is included in Appendix Section F of the Supporting Information.

Results

To contextualize our prediction results, we first examine between-group disparities in longevity for our focal 1910 birth cohort. These descriptive patterns highlight the well-known social gradients in mortality. We then assess how much of the individual-level variation in lifespan can be predicted using machine learning models trained on the same data.

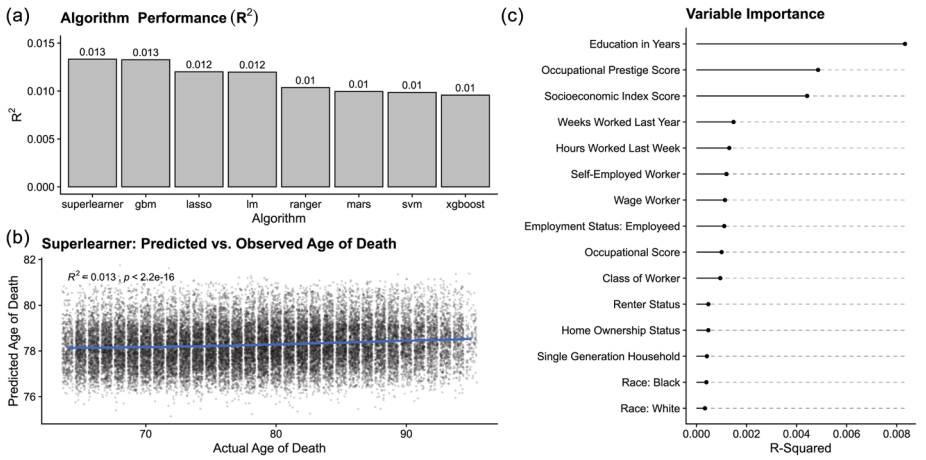
FIGURE 3 Comparison of within-group and between-group variation for a 5 percent random subsample of the birth cohort of 1910. (a) and (c) Group-level life expectancy across education and income groups. (b) and (d) Group-level life expectancy with individual lifespans overlaid, representing variation in individual outcomes within groups. Individual lifespans are horizontally jittered to improve visibility. Uncertainty bars show 95 percent confidence intervals (not visible in panels c and d because they are smaller than the plotted points)



Between- and within-group differences in longevity. As shown in Figure 3a, we observe a clear educational gradient in longevity, with higher educational attainment associated with higher life expectancy. We focus on educational attainment as it has a well-established relationship with mortality (Montez and Bisesti 2024) and is a common proxy for social class in the United States (Muller and Roehrkasse 2022; Pettit and Western 2004). A similar pattern is observed for income (Figure 3c), with higher wage and salary income associated with longer life expectancy. These between-group differences in life expectancy align in direction and approximate magnitude with other empirical studies (Halpern-Manners et al. 2020; Lleras-Muney, Price, and Yue 2020). These group-level differences reflect the between-group component of lifespan variation.

In Figures 3b and 3d, we present the same group-level estimates overlaid with individual-level death ages (represented by dots), which illustrate the within-group component of lifespan variation. The between-group differences in life expectancy are substantial: those with college degrees lived

FIGURE 4 (a) R^2 value for each machine learning algorithm. (b) Predicted versus observed values from the Superlearner algorithm. (c) Relative importance of the top 15 predictors.



nearly three years longer, on average, than those with only elementary education. However, these differences are small relative to the individual-level variation within groups. Many college graduates die before age 70, while many with only an elementary education live past age 90.

Individual-level predictions. We next assess how well our machine learning models predict individual lifespan. As shown in Figure 4a, none of the algorithms are able to accurately predict lifespan. The Superlearner algorithm—a weighted ensemble of the other algorithms—had the best out-of-sample performance in our holdout set ($R^2 = 0.013$). This low R^2 indicates that our best performing algorithm offers little improvement over simply predicting the sample mean age at death.

Figure 4b plots the correlation between the predicted age of death and the true, withheld age of death for the holdout sample. There is a very weak correlation between our predicted and observed age of death ($R = 0.114$). These results suggest that, even with core sociodemographic predictors and flexible algorithms, most of the variation in individual lifespan remains unexplained. Further, most of the predictions are narrowly concentrated between the ages of 76–82, indicating that the models largely fail to distinguish between individuals who die relatively early and those who live to older ages.

We next examine which predictors contribute most to the limited variation the model captures. Figure 4c plots variable importance for the top-performing Superlearner algorithm. Variable importance can be defined in many ways, each providing different insights into how an algorithm relies on a given variable for its predictions. We calculate variable

importance as the R^2 of each predictor from a univariate regression (for an alternative permutation-based variable importance, see Figure S5 in the Supporting Information). Education in years and occupational prestige score[10] were the two most important predictors, aligning with theoretical expectations (Galea et al. 2011; Montez and Bisesti 2024; Elo 2009).

We conducted several supplementary analyses. First, we examined predictability in a dataset that includes women, albeit with a shorter mortality observation window of 1988–2005. We found similarly low predictive accuracy for this sample ($R^2 = 0.012$). Although the shorter window limits comparability with the main results, this finding suggests that lifespan is likewise difficult to predict among women (Appendix Figure S1 of the Supporting Information). Second, we assessed predictive accuracy separately for each birth cohort from 1900 to 1920 (Appendix Figure S6 of the Supporting Information), finding that both earlier and later cohorts exhibited similarly low predictive accuracy. Finally, we evaluated predictive accuracy by race and socioeconomic status, finding that our models explained a smaller share of lifespan variance among lower education and lower income groups, as well as among Black Americans (Appendix Figure S7 of the Supporting Information). Given the large sample sizes, these likely reflect true greater underlying uncertainty in lifespan for these subgroups.

Discussion

The key methodological innovation of our study is that we make predictions within a cohort framework, focusing on a single birth cohort, rather than pooling individuals across ages. This design allows us to isolate the predictive power of sociodemographic characteristics net of age. We find that, despite large between-group differences in longevity, these sociodemographic characteristics explain less than 2 percent of the overall variation in lifespan. These results offer a new empirical lens into a well-established demographic fact: variation in mortality risk across groups is much smaller than the variation in lifespan within those groups (Vaupel 1988; Caswell 2009;2023). Theoretically, our findings center uncertainty as a fundamental feature of the human lifespan, highlighting that individual lifespan is only weakly determined by observable sociodemographic characteristics and shaped in large part by stochastic processes operating within structured social contexts.

These findings also echo classical results from the frailty modeling tradition. Vaupel (1988) empirically documented how the lifespans of parents and children are only weakly correlated, despite the intergenerational transmission of both genetic and environmental factors. Using simulation, Vaupel (1988) shows that even if frailty is directly inherited from parents, it still explains only 2–5 percent of the total variance in lifespan.[11] The limited explanatory power of socioeconomic factors, therefore, aligns with

the broader demographic insight that individual longevity is shaped only marginally by systematic determinants and largely by stochasticity.

Our results also add to a growing literature on the limits of prediction in other social domains (Salganik et al. 2020; Arpino, Le Moglie, and Mencarini 2022; Dressel and Farid 2018; Zheng and Cheng 2025). Comparisons across prediction exercises are inevitably limited by differences in data, features, and modeling approaches. Nonetheless, we situate our findings against those from other studies as illustrative points of reference. Rather than offering strict performance benchmarks, these results provide context for the scale of predictability we observed. For instance, Zheng and Cheng (2025) predict midlife socioeconomic status using 4,000 covariates from the stratification literature, finding a relatively high R^2 of 0.5. Compared with the life outcome prediction results in Salganik et al. (2020), our reported R^2 is lower than material hardship at age 15 ($R^2 = 0.23$), while primary caregiver layoff ($R^2 = 0.03$) represents a similarly low level of predictability. These comparisons reinforce that lifespan is a unique biosocial outcome shaped by both social and biological processes, with substantial stochasticity that limits predictive accuracy.

This low predictability can also be interpreted through the framework introduced by Lundberg et al. (2024), which decomposes prediction error for life outcomes into two components: learning error and irreducible error. Learning error reflects limitations in model fitting and sample size, while irreducible error captures within-group variance. In our setting, learning error is relatively small due to our large sample, while irreducible error is substantial: even within sociodemographic groups, variation in age of death remains high. This decomposition highlights that most variation in individual longevity remains unexplained by observed sociodemographic factors, reflecting an irreducible component of uncertainty.

On the one hand, the predictive power of observed covariates for longevity is limited; on the other hand, between-group inequalities in mortality remain substantial. Large differences in life expectancy across sociodemographic groups—by race, education, income, or place—can coexist with considerable within-group variation. Such disparities reflect structural inequality, policy-relevant gradients, and historically rooted disadvantage that remain central to mortality demography. Low predictive power should therefore not be equated with low structural inequality; rather, it reinforces mortality demography's focus on group-level outcomes and highlights the value of frameworks that integrate within- and between-group components to better understand lifespan inequality (Permanyer, Sasson, and Villavicencio 2023; Shi 2022).

Limited predictability clarifies how sociodemographic factors operate at the individual level. Sociodemographic characteristics may explain only a small share of individual-level lifespan variance while still generating meaningful differences in mortality risk and enabling discrimination

between individuals' relative survival prospects, especially when age is considered (Badolato et al. 2026). From this perspective, predictive analyses complement rather than replace traditional approaches by highlighting how structural gradients and individual-level uncertainty coexist. Prediction provides a quantitative lens for assessing the scale at which social determinants determine lifespan without implying deterministic life-course trajectories. Simply put, individual-level lifespan predictability is low because the between-group inequality is dominated by within-group variation (van Raalte et al. 2012; Vaupel 1988; Caswell 2023). This does not diminish the crucial importance of studying between-group disparities, which remain central and policy-relevant.

These findings can also be situated within the emerging field of uncertainty demography, which treats uncertainty as an inherent component of population processes deserving of study in its own right (Trinitapoli 2023). Our analysis extends this perspective to mortality by interpreting the predictability of individual lifespans as a measurable form of population-level uncertainty. Low predictability reflects the stochasticity inherent in mortality (uncertainty as an outcome), while that stochasticity is shaped by macro-level forces such as economic volatility, environmental shocks, epidemics, and what might be called life luck—forces that generate uncertainty as a cause of mortality variation. In this sense, uncertainty is not merely noise to be explained away by better covariates and models, but a defining demographic feature that structures both individual and collective experiences of life and death. More broadly, this stochasticity is a key feature not only of lifespan variation but also outcomes such as morbidity and lifetime reproductive output, where variance itself is a substantive object of analysis (Caswell 2023; van Daalen et al. 2022).

Building on this uncertainty demography perspective, our findings highlight how limited predictive performance can coexist with persistent structural inequalities. Rather than viewing limited predictability as a limitation, we interpret it as empirical evidence for a nature–nurture–chance framework, which calls for greater attention to the role of chance in shaping individual lifespans (Sasson 2025). This is operationalized here through high predictive uncertainty in cohort lifespan outcomes: prediction offers a complementary lens for assessing the balance between systematic social determinants and residual uncertainty. This aligns with evidence from studies of genetically identical twins, which find that substantial variability in longevity persists even when genetic and environmental factors are closely aligned (Finch and Kirkwood 2000).

This lifespan uncertainty has meaningful implications for lived experiences. Expectations about survival influence human capital investments (Sasson 2016), investment strategies (Abel 1985; Barro and Friedman 1977), and health and fertility decisions (Picone, Sloan, and Taylor 2004; Nettle 2010). Although our study does not directly examine behavioral

responses to uncertainty or perception thereof, the stratified patterns of predictability we document suggest that lifetime uncertainty is unevenly distributed across populations. Further, lifespan uncertainty may be especially pronounced and salient in violent or high-risk contexts, where mortality risks both shorten lives and increase unpredictability in the timing of death, complicating long-term planning and decision-making (Aburto et al. 2023).

Several limitations and opportunities for future research warrant discussion. Our main analysis is restricted to men born in 1910, observed in the 1940 Census, and dying between ages 65 and 95. While this design enables a cohort-based approach to lifespan prediction, it excludes early deaths before age 65; nevertheless, approximately 65 percent of this cohort died within the observed window (Breen and Osborne 2022). Our predictors are measured cross-sectionally in early adulthood at age 29 or 30, and certain predictors, such as income, occupation, or marital status, may change over the life course. As a result, we cannot assess how life course trajectories or covariates measured at older ages might alter predictability, even though such measures could plausibly carry stronger mortality signals (e.g., marital dissolution in later life). Longitudinal life-course data may improve predictive performance relative to our baseline; however, even substantial improvement—for example, a doubling or tripling of explanatory power—would still imply relatively low overall predictability. Moreover, several key predictors, including educational attainment, are largely fixed by early adulthood for this cohort.

By design, we focused on sociodemographic variables that capture systemic inequality. Biological and behavioral predictors are not the primary focus of this study, which centers on the social structuring of mortality. Such predictors do, however, overlap with the sociodemographic characteristics considered here, and future work could explicitly consider these predictors.

Our primary analysis focuses on the cohort of 1910 and may not generalize to other more recent birth cohorts. Replication efforts with cohorts born between 1900 and 1920 revealed similarly low predictability (R^2 values ranging from 0.5 percent to 1.4 percent). Our main analysis is also focused on men due to data constraints: surname changes for women at marriage precluded their linkage between the 1940 Census and mortality records. However, in Appendix Section C of the Supporting Information, we replicate our analysis on a similar linked file that includes a shorter mortality observation window but includes both genders. We find similarly low predictability in this sample with both genders ($R^2 = 0.012$).

Taken together, our findings highlight the limits of sociodemographic characteristics in structuring one of life's most consequential outcomes: how long one will live. Structural inequality generates large and meaningful differences in life expectancy across groups, yet these same characteristics explain little of the variation in lifespan among individuals. This underscores a central tension: while between-group disparities are substantial and socially

significant, the within-group variation dominates individual lifespans. This demonstrates the fundamentally nondeterministic nature of how structural inequality shapes individual lifespan.

Acknowledgments

For helpful discussions and feedback, we thank Hal Caswell, Jenn Dowd, Aashish Gupta, Dennis Feehan, Joshua R. Goldstein, Ridhi Kashyap, Jenna Nobles, Patricia McManus, Chris Muller, Michelle Niemann, Mathew Salganik, Alyson Van Raalte, Ken Wachter, Elizabeth Wrigley-Field, participants of the Berkeley CenSoc working group, participants in the Oxford Health Inequalities Reading Group, participants in the PAA 2023 "Socioeconomic Inequalities in Mortality" session and participants in the ASA 2022 "Computational Sociology: Methods and Applications" session. C.F.B. was supported by the National Institute of Aging T32-AG000246. Replication code is available from https://osf.io/fazsj/?view only = cbfea9a08a684a48b1c97d1e5f8da967. The research reported in this publication was supported by the National Institute on Aging (NIA) (R01AG05894) and infrastructure grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (P2CHD042849) and the NIA (P30AG066614).

Notes

1 Trinitapoli (2023) frame uncertainty demography as treating uncertainty not as the amount of error in an estimate but as a constitutive feature of social and demographic life. It examines how populations experience, reproduce, and navigate uncertainty as a social fact.

2 This stochastic variation may reflect truly random processes or latent individual-level determinants. What appears as "luck" in lifespan may instead reflect fine-grained variation in exposures, behaviors, genetics, or life events that are unmeasured. This differs from variance calculated from the rates estimated for groups using life-table or Markov formulations, which assume homogeneity within groups by construction (Caswell 2023).

3 A strength of the predictive approach is its ability to identify the strongest individual-level predictors of mortality. Goldman, Glei, and Weinstein (2016);2017) found that self-rated health, mobility limitations, and difficulties with instrumental activities of daily living consistently outperformed clinical measures like obesity or diabetes in predicting survival. Puterman et al. (2020) found the most important predictors of longevity included smoking behavior, alcohol abuse, and history of divorce.

4 For example, Ottenhoff et al. (2021) predicted 21-day mortality for hospitalized COVID-19 patients using sociodemographic and clinical data. While these models perform reasonably well for short time horizons, they rely on detailed clinical data not typically available in population datasets.

5 We cannot rule out the possibility that model performance will improve in the future. As Yan and Rahal (2025) notes, the predictability of social systems remains uncertain: data and algorithms continue to improve, and claims of "unpredictability" are context dependent.

6 Prior work has examined cohort patterns in lifespan variance and frailty in several European countries for historical cohorts, finding that only a small amount of

lifespan variance is due to frailty (Hartemink, Missov, and Caswell 2017).

7 Approximately 9 percent of records had at least one missing value. Imputing these missing values using multiple imputation, rather than dropping cases with missing values, produced comparable but slightly lower predictive performance.

8 The motivation behind the ensemble Superlearner approach is that a weighted combination of different algorithms generally outperforms any single algorithm. The ensemble Superlearner algorithm selects the optimal weighted combination using a cross-validation procedure within the training set to minimize overfitting risk (Phillips et al. 2023). For our ensemble Superlearner, we include a set of widely used machine learning algorithms: random forests, linear regression, gradient boosting machines, lasso regression, extreme gradient boosting machines, and support vector machines.

9 We chose a classic train/test split over *k*-fold cross-validation due to computational limitations of our large sample. Given the large sample size, a single holdout set is sufficient for reliable out-of-sample evaluation while avoiding the prohibitive computational cost of repeatedly reestimating ensemble models required by *k*-fold cross-validation.

10 We use the Siegel occupational prestige score, which calculates an occupational prestige score for each occupation based on perceived status from a survey of the general population.

11 In this exercise by Vaupel (1988), frailty is represented by a single parameter that aggregates all mortality risk. In the simulations, each individual receives a parameter value, which functions as a stand-in for their relative risk in a proportional hazards model.

References

Abel, A. B. 1985. "Precautionary Saving and Accidental Bequests." *The American Economic Review* 75 (4): 777–791.

Aburto, J. M., di Lego, V., Riffe, T., Kashyap, R., van Raalte, A., and Torrisi, O. 2023. "A Global Assessment of the Impact of Violence on Lifetime Uncertainty." *Science Advances* 9 (5): eadd9038.

Alexander, M. 2018. "Deaths without Denominators: Using a Matched Dataset to Study Mortality Patterns in the United States." Preprint, SocArXiv q79ye. Center for Open Science.

Arpino, B., Le Moglie, M., and Mencarini, L. 2022. "What Tears Couples Apart: A Machine Learning Analysis of Union Dissolution in Germany." *Demography* 59 (1): 161–186.

Badolato, L., Decter-Frain, A., Irons, N. J., Miranda, M. L., Walk, E., Zhalieva, E., Alexander, M., Basellini, U., and Zagheni, E. 2026. "The Limits of Predicting Individual-Level Longevity: Insights from the U.S. Health and Retirement Study." *Demography* 63 (1): 351–374.

Barro, R. J., and Friedman, J. W. 1977. "On Uncertain Lifetimes." *Journal of Political Economy* 85 (4): 843–849.

Breen, C., and Osborne, M. 2022. "An Assessment of CenSoc Match Quality." Preprint, SocArXiv bjm5md_ v1. CenSoc project.

Breen, C. F., Osborne, M., and Goldstein, J. R. 2023. "CenSoc: Public Linked Administrative Mortality Records for Individual-level Research." *Scientific Data* 10(1): 802.

Caswell, H. 2009. "Stage, Age and Individual Stochasticity in Demography." *Oikos* 118 (12): 1763–1782.

Caswell, H. 2023. "The Contributions of Stochastic Demography and Social Inequality to Lifespan Variability." *Demographic Research* 49: 309–354.

Caswell, H., and Van Daalen, S. F. 2025. "Inequality, Heterogeneity, and Chance: Multiple Factors and Their Interactions." *Vienna Yearbook of Population Research* 23: 129–155.

Chetty, R., and Hendren, N. 2018. "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*." *The Quarterly Journal of Economics* 133 (3): 1107–1162.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., and Cutler, D. 2016. "The Association between Income and Life Expectancy in the United States, 2001-2014." *JAMA* 315 (16): 1750.

Cutler, D., Deaton, A., and Lleras-Muney, A. 2006. "The Determinants of Mortality." *Journal of Economic Perspectives* 20 (3): 97–120.

Dannefer, D. 2003. "Cumulative Advantage/Disadvantage and the Life Course: Cross-Fertilizing Age and Social Science Theory." *The Journals of Gerontology: Series B* 58 (6): S327–S337.

Dowd, J. B., Doniec, K., Zhang, L., and Tilstra, A. 2024. "US Exceptionalism? International Trends in Midlife Mortality." *International Journal of Epidemiology* 53 (2): dyae024.

Dowd, J. B., Polizzi, A., and Tilstra, A. M. 2025. "Progress Stalled? The Uncertain Future of Mortality in High-Income Countries." *Population and Development Review* 51 (1): 257–293.

Dressel, J., and Farid, H. 2018. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Science Advances* 4 (1): eaao5580.

Einav, L., Finkelstein, A., Mullainathan, S., and Obermeyer, Z. 2018. "Predictive Modeling of U.S. Health Care Spending in Late Life." *Science* 360 (6396): 1462–1465.

Elo, I. T. 2009. "Social Class Differentials in Health and Mortality: Patterns and Explanations in Comparative Perspective." *Annual Review of Sociology* 35: 553–572.

Elo, I. T., Turra, C. M., Kestenbaum, B., and Ferguson, B. R. 2004. "Mortality Among Elderly Hispanics in the United States: Past Evidence and New Results." *Demography* 41 (1): 109–128.

Feigenbaum, J. J., Muller, C., and Wrigley-Field, E. 2019. "Regional and Racial Inequality in Infectious Disease Mortality in U.S. Cities, 1900–1948." *Demography* 56 (4): 1371–1388.

Finch, C. E., and Kirkwood, T. B. L. 2000. *Chance, Development and Aging*. Oxford: Oxford University Press.

Fletcher, J., and Noghanibehambari, H. 2024. "The Effects of Education on Mortality: Evidence Using College Expansions." *Health Economics* 33 (3): 541–575.

Galea, S., Tracy, M., Hoggatt, K. J., Dimaggio, C., and Karpati, A. 2011. "Estimated Deaths Attributable to Social Factors in the United States." *American Journal of Public Health* 101 (8): 1456–1465.

Goldman, N., Glei, D. A., and Weinstein, M. 2016. "What Matters Most for Predicting Survival? A Multinational Population-Based Cohort Study." *PLoS ONE* 11(7): e0159273.

Goldman, N., Glei, D. A., and Weinstein, M. 2017. "The Best Predictors of Survival: Do They Vary by Age, Sex, and Race?" *Population and Development Review* 43 (3): 541–560.

Goldstein, J. R., Alexander, M., Breen, C., Miranda González, A., Menares, F., Osborne, M., Snyder, M., and Yildirim, U. 2021. "Censoc Project." CenSoc Mortality File: Version 2.0. Berkeley: University of California.

Gutin, I., and Hummer, R. A. 2021. "Social Inequality and the Future of US Life Expectancy." *Annual Review of Sociology* 47: 501–520.

Halpern-Manners, A., Helgertz, J., Warren, J. R., and Roberts, E. 2020. "The Effects of Education on Mortality: Evidence From Linked U.S. Census and Administrative Mortality Data." *Demography* 57 (4): 1513–1541.

Hartemink, N., Missov, T. I., and Caswell, H. 2017. "Stochasticity, Heterogeneity, and Variance in Longevity in Human Populations." *Theoretical Population Biology* 114: 107–116.

Hayward, M. D., and Gorman, B. K. 2004. "The Long Arm of Childhood: The Influence of Early-Life Social Conditions on Men's Mortality." *Demography* 41 (1): 87–107.

Hayward, M. D., and Heron, M. 1999. "Racial Inequality in Active Life among Adult Americans." *Demography* 36 (1): 77–91.

Hill, M. E. 2001. "The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages." *Social Security Bulletin* 64 (1): 45–51.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., and Yarkoni, T. 2021. "Integrating Explanation and Prediction in Computational Social Science." *Nature* 595 (7866): 181–188.

Hummer, R. A. 2000. "Adult Mortality Differentials among Hispanic Subgroups and Non-Hispanic Whites." *Social Science Quarterly* 81 (1): 459–476.

Hummer, R. A., Benjamins, M. R., and Rogers, R. G. 2004. "Racial and Ethnic Disparities in Health and Mortality among the US Elderly Population." In *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*, 53–94. Washington, DC: National Academy Press.

Kapoor, S., Cantrell, E. M., Peng, K., Pham, T. H., Bail, C. A., Gundersen, O. E., Hofman, J. M., Hullman, J., Lones, M. A., Malik, M. M., Nanayakkara, P., Poldrack, R. A., Raji, I. D., Roberts, M., Salganik, M. J., Serra-Garcia, M., Stewart, B. M., Vandewiele, G., and Narayanan, A. 2024. "REFORMS: Consensus-Based Recommendations for Machine-Learning-Based Science." *Science Advances* 10 (18): eadk3452.

Kashyap, R. 2021. "Has Demography Witnessed a Data Revolution? Promises and Pitfalls of a Changing Data Ecosystem."*Population Studies* 75 (sup1): 47–75.

Lariscy, J. T., Hummer, R. A., and Hayward, M. D. 2015. "Hispanic Older Adult Mortality in the United States: New Estimates and an Assessment of Factors Shaping the Hispanic Paradox." *Demography* 52 (1): 1–14.

Link, B. G., and Phelan, J. 1995. "Social Conditions as Fundamental Causes of Disease." *Journal of Health and Social Behavior* 35: 80–94.

Lleras-Muney, A., Price, J., and Yue, D. 2020. "The Association between Educational Attainment and Longevity Using Individual Level Data from the 1940 Census." NBER Working Paper Series.

Lundberg, I., Brand, J. E., and Jeon, N. 2022. "Researcher Reasoning Meets Computational Capacity: Machine Learning for Social Science." *Social Science Research* 108: 102807.

Lundberg, I., Brown-Weinstock, R., Clampet-Lundquist, S., Pachman, S., Nelson, T. J., Yang, V., Edin, K., and Salganik, M. J. 2024. "The Origins of Unpredictability in Life Outcome Prediction Tasks." *Proceedings of the National Academy of Sciences* 121 (24): e2322973121.

Montez, J. K., and Bisesti, E. M. 2024. "Widening Educational Disparities in Health and Longevity." *Annual Review of Sociology* 50: 547–564.

Montez, J. K., Harward, M. D., and Wolf, D. A. 2017. "Do U.S. States' Socioeconomic and Policy Contexts Shape Differences in Adult Disability?" *Social Science & Medicine*, 178, 115–126.

Montez, J. K., Hummer, R. A., and Hayward, M. D. 2012. "Educational Attainment and Adult Mortality in the United States: A Systematic Analysis of Functional Form." *Demography* 49 (1): 315–336.

Muller, C., and Roehrkasse, A. F. 2022. "Racial and Class Inequality in US Incarceration in the Early Twenty-First Century." *Social Forces* 101, 2803–28.

Nettle, D. 2010. "Dying Young and Living Fast: Variation in Life History across English Neighborhoods." *Behavioral Ecology* 21 (2): 387–395.

Ottenhoff, M. C., Ramos, L. A., Potters, W., Janssen, M. L. F., Hubers, D., Hu, S., Fridgeirsson, E. A., Piña-Fuentes, D., Thomas, R., van der Horst, I. C. C., Herff, C., Kubben, P., Elbers, P. W. G., Marquering, H. A., Welling, M., Simsek, S., de Kruif, M. D., Dormans, T., Fleuren, L. M., Schinkel, M., Noordzij, P. G., van den Bergh, J. P., Wyers, C. E., Buis, D. T. B., Wiersinga, W. J., van den Hout, E. H. C., Reidinga, A. C., Rusch, D., Sigaloff, K. C. E., Douma, R. A., de Haan, L., van den Oever, N. C. G., Rennenberg, R. J. M. W., van Wingen, G. A., Aries, M. J. H., and Beudel, M. 2021. "Predicting Mortality of Individual Patients with COVID-19: A Multicentre Dutch Cohort." *BMJ Open* 11 (7): e047347.

Permanyer, I., Sasson, I., and Villavicencio, F. 2023. "Group- and Individual-Based Approaches to Health Inequality: Towards an Integration." *Journal of the Royal Statistical Society Series A: Statistics in Society* 186 (2): 217–240.

Permanyer, I., Spijker, J., Blanes, A., and Renteria, E. 2018. "Longevity and Lifespan Variation by Educational Attainment in Spain: 1960-2015." *Demography* 55 (6): 2045–2070.

Pettit, B., and Western, B. 2004. "Mass Imprisonment and the Life Course: Race and Class Inequality in U.S. Incarceration." *American Sociological Review* 69 (2): 151–169.

Phillips, R. V., van der Laan, M. J., Lee, H., and Gruber, S. 2023. "Practical Considerations for Specifying a Super Learner." *International Journal of Epidemiology* 52 (4): 1276–1285.

Picone, G., Sloan, F., and Taylor, D. 2004. "Effects of Risk and Time Preference and Expected Longevity on Demand for Medical Tests." *Journal of Risk and Uncertainty* 28 (1): 39–53.

Preston, S. H., and Elo, I. T. 1995. "Are Educational Differentials in Adult Mortality Increasing in the United States?" *Journal of Aging and Health* 7 (4): 476–496.

Puterman, E., Weiss, J., Hives, B. A., Gemmill, A., Karasek, D., Mendes, W. B. & Rehkopf, D. H. 2020. "Predicting Mortality from 57 Economic, Behavioral, Social, and Psychological Factors." *Proceedings of the National Academy of Sciences* 117 (28): 16273–16282.

Rose, S. 2013. "Mortality Risk Score Prediction in an Elderly Population Using Machine Learning." *American Journal of Epidemiology* 177 (5): 443–452.

Ruggles, S., Flood, S., Goeken, R., Grover, J., Meyer, E., Pacas, J., and Sobek, M. 2020. IPUMS USA: Version 10.0 [Dataset]. Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., Morgan, A. C., Pentland, A., Polimis, K., Raes, L., Rigobon, D. E., Roberts, C. V., Stanescu, D. M., Suhara, Y., Usmani, A., Wang, E. H., Adem, M., Alhajri, A., AlShebli, B., Amin, R., Amos, R. B., Argyle, L. P., Baer-Bositis, L., Büchi, M., Chung, B.-R., Eggert, W., Faletto, G., Fan, Z., Freese, J., Gadgil, T., Gagné, J., Gao, Y., Halpern-Manners, A., Hashim, S. P., Hausen, S., He, G., Higuera, K., Hogan, B., Horwitz, I. M., Hummel, L. M., Jain, N., Jin, K., Jurgens, D., Kaminski, P., Karapetyan, A., Kim, E. H., Leizman, B., Liu, N., Möser, M., Mack, A. E., Mahajan, M., Mandell, N., Marahrens, H., Mercado-Garcia, D., Mocz, V., Mueller-Gastell, K., Musse, A., Niu, Q., Nowak, W., Omidvar, H., Or, A., Ouyang, K., Pinto, K. M., Porter, E., Porter, K. E., Qian, C., Rauf, T., Sargsyan, A., Schaffner, T., Schnabel, L., Schonfeld, B., Sender, B., Tang, J. D., Tsurkov, E., van Loon, A., Varol, O., Wang, X., Wang, Z., Wang, J., Wang, F., Weissman, S., Whitaker, K., Wolters, M. K., Woon, W. L., Wu, J., Wu, C., Yang, K., Yin, J., Zhao, B., Zhu, C., Brooks-Gunn, J., Engelhardt, B. E., Hardt, M., Knox, D., Levy, K., Narayanan, A., Stewart, B. M., Watts, D. J., and McLanahan, S. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117 (15): 8398–8403.

Sasson, I. 2016. "Trends in Life Expectancy and Lifespan Variation by Educational Attainment: United States, 1990-2010." *Demography* 53 (2): 269–293.

Sasson, I. 2025. "A New Research Agenda for Social Inequalities in Mortality: Challenges and Open Questions." *Population and Development Review* 51 (1): 323–360.

Savcisens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L. H., Lilleholt, L., Rogers, A., Zettler, I., and Lehmann, S. 2024. "Using Sequences of Life-Events to Predict Human Lives." *Nature Computational Science* 4 (1): 43–56.

Schwandt, H., Currie, J., Bär, M., Banks, J., Bertoli, P., Bütikofer, A., Cattan, S., Chao, B. Z.-Y., Costa, C., González, L., Grembi, V., Huttunen, K., Karadakic, R., Kraftman, L., Krutikova, S., Lombardi, S., Redler, P., Riumallo-Herl, C., Rodríguez-González, A., Salvanes, K. G., Santana, P., Thuilliez, J., van Doorslaer, E., Van Ourti, T., Winter, J. K., Wouterse, B., and Wuppermann, A. 2021. "Inequality in Mortality between Black and White Americans by Age, Place, and Cause and in Comparison to Europe, 1990 to 2018." *Proceedings of the National Academy of Sciences* 118 (40): e2104684118.

Seaman, R., Riffe, T., and Caswell, H. 2019. "Changing Contribution of Area-Level Deprivation to Total Variance in Age at Death: A Population-Based Decomposition Analysis." *BMJ Open* 9 (3): e024952.

Shi, J. 2022. "Decomposing Lifespan Variance: A Distributional Approach." Preprint, SocArXiv. November 28, 2025. Available from: https://doi.org/10.31235/osf.io/zd79q

Snyder, R. E., and Ellner, S. P. 2018. "Pluck or Luck: Does Trait Variation or Chance Drive Variation in Lifetime Reproductive Success?" *The American Naturalist* 191 (4): E90–E107.

Steiner, U. K., Tuljapurkar, S., and Orzack, S. H. 2010. "Dynamic Heterogeneity and Life History Variability in the Kittiwake." *Journal of Animal Ecology* 79 (2): 436–444.

Trinitapoli, J. 2023. *An Epidemic of Uncertainty: Navigating HIV and Young Adulthood in Malawi*. Chicago: University of Chicago Press.

van Daalen, S. F., Hernández, C. M., Caswell, H., Neubert, M. G., and Gribble, K. E. 2022. "The Contributions of Maternal Age Heterogeneity to Variance in Lifetime Reproductive Output." *The American Naturalist* 199 (5): 603–616.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. 2007. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1): 5.

van Raalte, A. A. 2021. "What Have We Learned about Mortality Patterns over the Past 25 Years?" *Population Studies* 75(sup1): 105–132.

van Raalte, A. A., Kunst, A. E., Lundberg, O., Leinsalu, M., Martikainen, P., Artnik, B., Deboosere, P., Stirbu, I., Wojtyniak, B., and Mackenbach, J. P. 2012. "The Contribution of Educational Inequalities to Lifespan Variation." *Population Health Metrics* 10 (1): 3.

Vaupel, J. W. 1988. "Inherited Frailty and Longevity." *Demography* 25 (2): 277–287.

Vaupel, J. W., Manton, K. G., and Stallard, E. 1979. "The Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality." *Demography* 16 (3): 439–454.

Wrigley-Field, E. 2020. "Multidimensional Mortality Selection: Why Individual Dimensions of Frailty Don't Act Like Frailty." *Demography* 57 (2): 747–777.

Wrigley-Field, E. 2025. "Three Ways of Looking at Black–White Mortality Differences in the United States." *Annual Review of Sociology* 51 (1): 311–333.

Yan, J., and Rahal, C. 2025. "On the Unknowable Limits to Prediction." *Nature Computational Science* 5 (3): 188–190.

Zheng, H., and Cheng, S. 2025. "Social Rigidity across and Within Generations: A Predictive Approach." *Sociological Methods, and Research* 54 (4): 1683–1725.

Supplemental Materials

Estimating death rates in complex humanitarian emergencies using the network survival method

Casey F. Breen and Nathan Seltzer

A List of predictors

All predictors for our analyses come from the complete-count 1940 Census. The full set of predictors is shown below in Table A1. We recoded categorical variables into binary (0/1) indicators and standardized continuous predictors to a common scale, with a mean of 0 and a standard deviation of 1.

Variable	Description
Education (years)	Educational attainment in years.
Wage and salary income	Wage and salary income for employees, topcoded at $5,001
Socioeconomic Index Score	Duncan Socioeconomic Index, a composite score reflecting occupational status and earnings
Occupation Income Score	Occupational income score based on median income of all people with a given occupation in 1950.
Household size	Number of persons living in a household, topcoded at 20.
Occupation Prestige Score	Occupational prestige score, based on Siegel's method.
Race: White	Binary variable for whether an individual was classified as White.
Race: Black	Binary variable for whether an individual was classified as Black.
Race: Other	Binary variable for whether an individual was not classified as Black or White.
Metro	Binary variable indicating lives in a metro area.
City	Binary variable indicating whether an individual lived in a city.
Suburb	Binary variable indicating whether an individual lived in a suburb.

Marital Status: Married	Binary variable indicating whether an individual is married.
Marital Status: Divorced	Binary variable indicating whether an individual is divorced.
Marital Status: Widowed	Binary variable indicating whether an individual is widowed.
Marital Status: Single	Binary variable indicating whether an individual is single.
Self-Employed	Binary indicator on whether an individual is self-employed.
Wage Worker	Binary indicator on whether an individual is a wage and salary workers.
Lives with Mother	Binary indicator on whether someone lives in the same household as their mother
Migration: No movement	Binary variable indicating no movement from house in the last 5 years.
Migration: Internal	Binary variable for individuals who moved states in the last 5 years.
Migration International: International	Binary variable for individuals who moved countries in the last 5 years.
Urban	Binary variable distinguishing urban versus rural place of residence.
Labor force Participation	Binary variable indicating labor force participation status.
Farm	Binary variable indicating if the residence is on a farm.
Homeowner	Binary variable indicating homeownership vs. renter status.
Household: Single Generation	Binary variable indicating a single generation household.
Household: Second Generation	Binary variable indicating a two-generation household.
Household: Third Generation	Binary variable indicating a three-generation household.
Employment Status	Binary employment status variable for employed individuals.
State	State of residence in 1940.

Table A1: Variables used for prediction

B Machine Learning Approach

For most prediction tasks, it is impossible to know in advance which predictive algorithm will have the best performance. To overcome this, we use Superlearning. Superlearning—also known as weighted ensembling or model stacking—is a statistical technique for prediction that combines many machine learning algorithms into a single algorithm (Van der Laan, Polley and Hubbard, 2007). The ensemble Superlearner algorithm finds the best weighted combination of algorithms using a k-fold cross-validation procedure to minimize cross-validated risk (Van der Laan, Polley and Hubbard, 2007). In the Superlearning framework, a metalearner is used to combine the predictions of the discrete algorithms.[13] This metalearner is trained exclusively on the training data, using the outputs of the base models as input features.

The motivation behind Superlearning is that a weighted combination of many different algorithms may outperform any single algorithm by smoothing out the limitations of any specific model (Phillips et al., 2023). This avoids the issue of picking an algorithm a priori by selecting the best weighted combination of algorithms using a pre-defined set of decision rules that minimize cross-validated mean squared error.

We fit the Superlearner algorithm on our training data using the following implementation:

1. Split the data into $k = 10$ different folds (partitions) for k-fold cross-validation.

2. Using the first fold as the holdout data, fit each discrete algorithm on the 9 other folds. Make predictions on the holdout fold using each algorithm. Repeat this process for

[13] Here, we use a non-negative least squares regression model as our metalearner.

each fold, using a different fold as the holdout data.

3. Choose a loss function (e.g., mean squared error) and compute a metalearner. The metalearner is a principled approach to combining multiple machine learning algorithms by fitting a regression of outcome variables on the predicted variables to minimize the cross-validated risk of the set of potential weighted combinations.

4. Fit each of the algorithms on the full data.

In addition to asymptotically producing a weighted algorithm that theoretically performs as well or better than the top algorithm, the Superlearner has the added advantage of training many machine learning algorithms at the same time. This allows us to assess the performance of many different machine learning algorithms with an a priori specified evaluation criterion: minimum cross-validated mean squared error. This is a conservative criterion that prioritizes minimizing large predictive errors.

Algorithm	Description	Cross-validated Risk	Superlearner Coefficient
gbm	Gradient boosting machines	63.61	0.47
lm	Linear model	63.63	0.09
xgboost	Extreme gradient boosting	63.89	0.07
ranger	Random forest regression	63.87	0.14
svm	Support vector machine	63.76	0.23
lasso	Lasso regression	63.88	0.00
mars	Multivariate adaptive regression splines	63.63	0.00
mean	Arithmetic mean	64.43	0.00
Superlearner	Ensemble Superlearner	63.59	—

Table A2: Full set of algorithms included in the Superlearner ensemble model for cohort analysis. The cross-validated risk refers to the risk calculated by the Superlearner, the cross-validated mean squared error. The coefficient gives the total weight (contribution) of each algorithm toward the full Superlearner ensemble model.

Table A2 shows the full set of machine learning algorithms used in our Superlearner.

The cross-validated risk is lowest for the gradient boosting machines and highest for the lasso regression and extreme gradient boosting. The Superlearner algorithm most heavily weighted the gradient boosting machines algorithm and the support vector machine algorithm.

B.1 Assessing model performance

Our primary metric for assessing predictability is $R^2_{holdout}$, defined as:

$$R^2_{holdout} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

where y_i is the observed age of death in the holdout dataset, $\hat{y}_i$ is the predicted age of death in the holdout dataset, $\bar{y}$ is the mean observed age of death in the holdout dataset, and n is the number of observations. This is a continuous metric that ranges from 0, which represents predictions no better than the mean of the holdout dataset, to 1, indicating perfect predictions. We use this metric to facilitate comparisons with other studies (Salganik et al., 2020; Puterman et al., 2020; Zheng and Cheng, 2025).

As a second measure of predictive performance, we calculate the mean absolute error, the average of the absolute differences between the predicted and true lifespan in our holdout sample. The mean absolute error is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

where y_i is the observed age of death for individual i in the holdout dataset, $\hat{y}_i$ is the predicted age of death for individual i in the holdout dataset, and n is the total number of observations.

This intuitive metric provides insight into how close, on average, our predicted lifespans are to the true lifespans.

C Prediction for women

For our main analysis, we use the CenSoc-DMF, which includes deaths for the window of 1975–2005 but is restricted to men. In this supplementary analysis, we replicate our main prediction exercise using data from the CenSoc-Numident, which links the 1940 Census with mortality records from the Social Security Numident (Goldstein et al., 2021). In contrast with the CenSoc-DMF, the CenSoc-Numident only includes deaths for a shorter window, between 1988–2005. However, the CenSoc-Numident file includes women, allowing us to replicate our analysis on a dataset that includes women.

Results from this replication are shown in Figure A1. The findings are comparable to our CenSoc-DMF results: all models have very low predictive accuracy. The top-performing Superlearner model explains only 1.2% of the variation in age at death.

D Period perspective

While our main analysis predicts age at death from a cohort perspective, we also conduct a period-based analysis to test the predictability of shorter-term mortality risk of individuals of different ages. Specifically, we predict 5-year mortality using a random sample of 100,000 men aged 54–95 in 1975. This design resembles a cross-sectional survey with follow-up, allowing us to include age as a predictor.

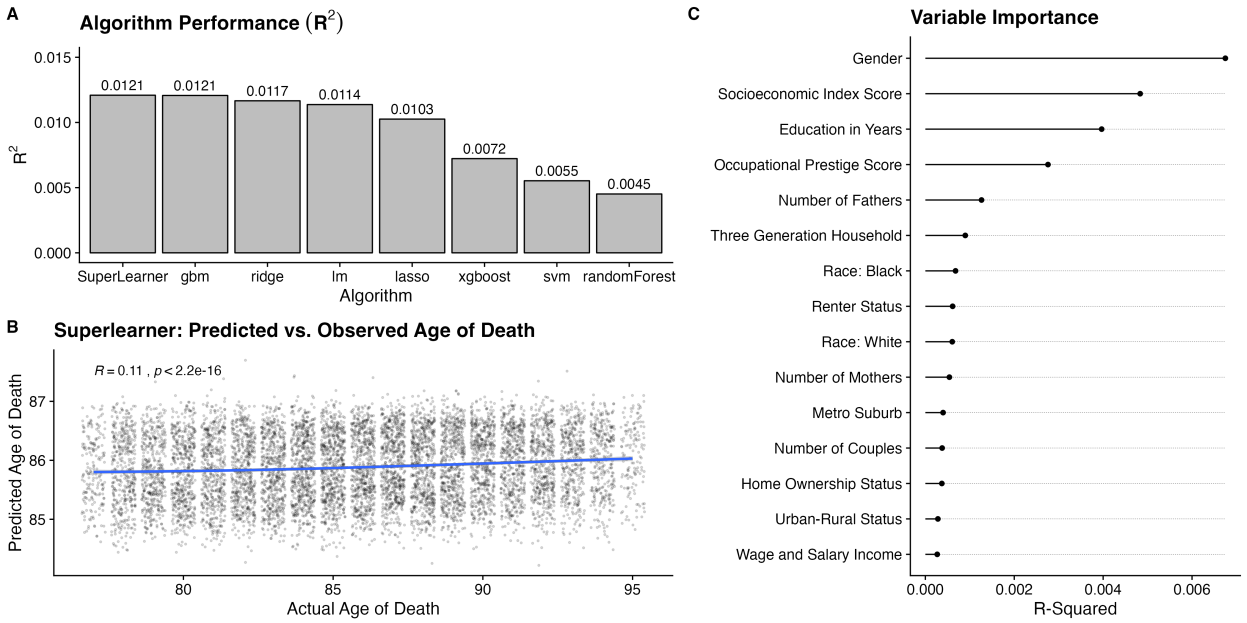


Figure A1: *(A)* R^2 value for each machine learning algorithm. *(B)* Predicted vs. observed values from Superlearner algorithm. *(C)* Relative importance of the top 15 predictors. The CenSoc-Numident cohort of 1910 includes 75,238 individuals (20,167 women and 55,071 men).

We use the same set of sociodemographic predictors as in the cohort analysis with the addition of age. Instead of predicting lifespan, we model a binary outcome: whether an individual dies in a 5-year window between 1975 and 1979. We fit a Superlearner ensemble using algorithms appropriate for classification tasks: logistic regression, elastic net, ridge regression, lasso, random forest, and extreme gradient boosting.

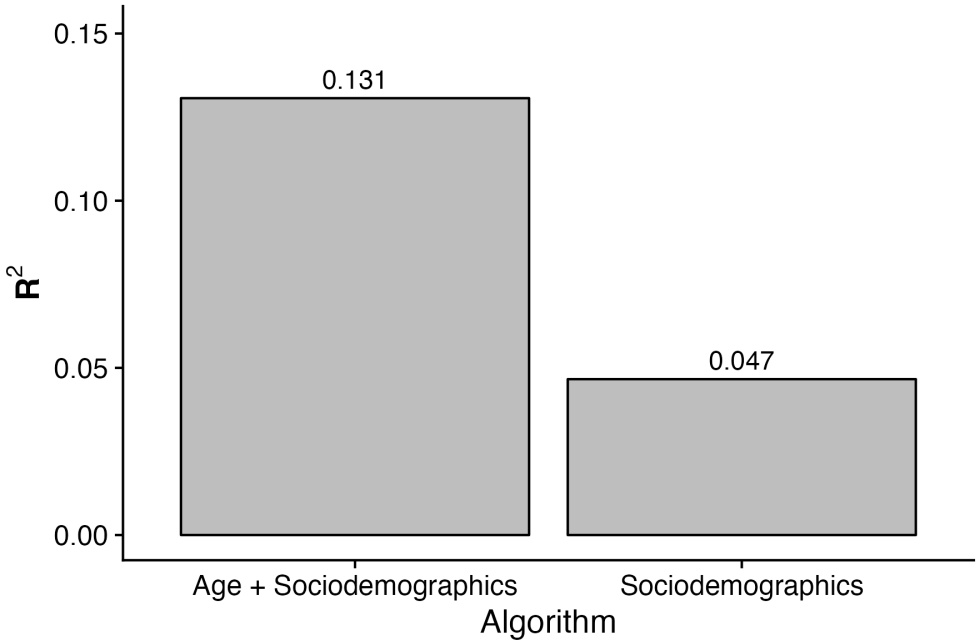


Figure A2: Period analysis results. R^2 for model using age and sociodemographic predictors and models using sociodemographic predictors alone.

Figure A2 shows that the model using both age and sociodemographic predictors achieves modest accuracy ($R^2 = 0.131$). A model trained exclusively on sociodemographics without age performs substantially worse ($R^2 = 0.047$). This suggests limited incremental predictive value from early-life sociodemographic factors. This analysis also reinforces that predicting mortality in the next 5 years results in higher predictive accuracy than using the same characteristics to predict lifespan in a cohort framework.

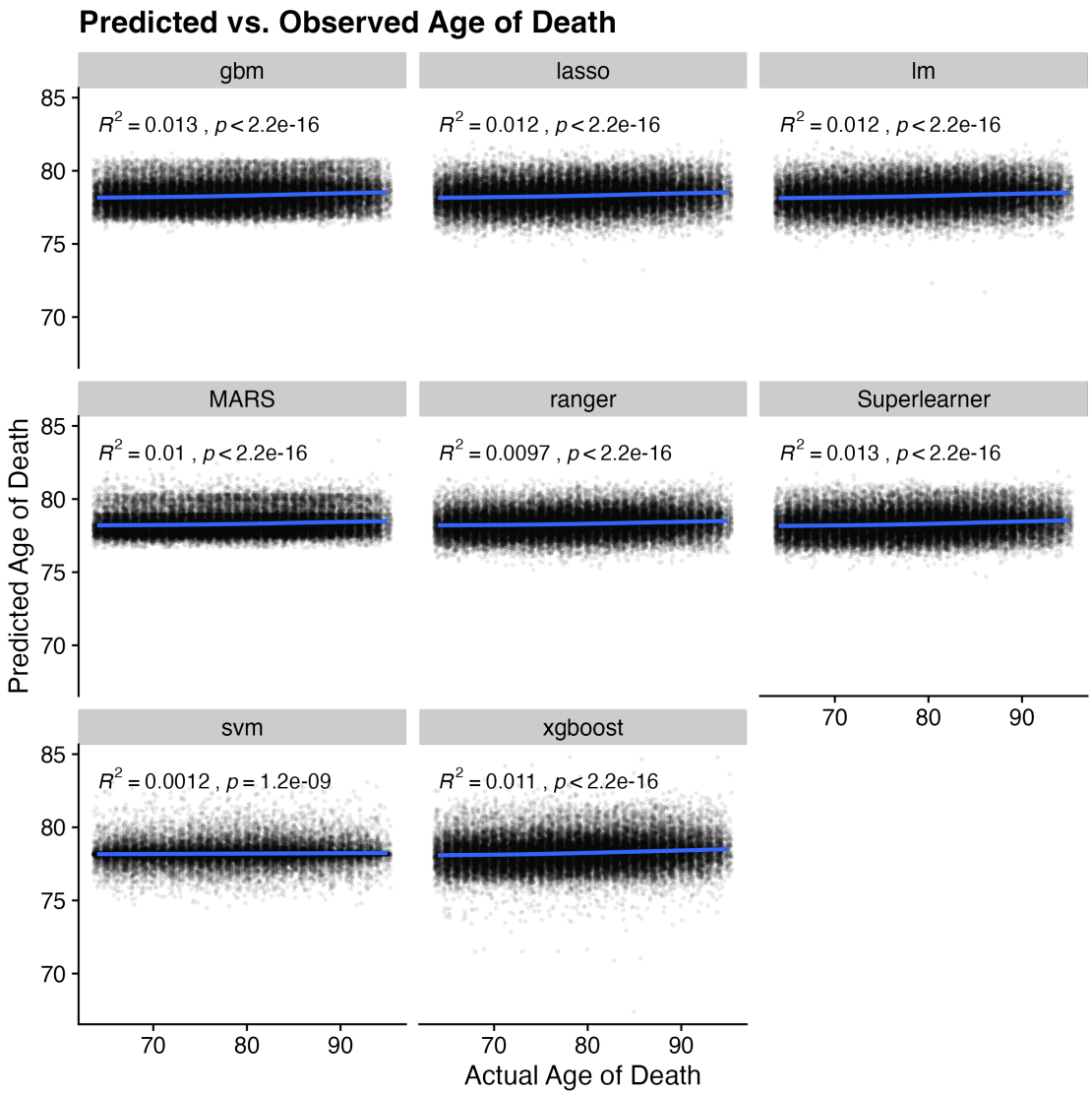


Figure A3: Scatterplots showing predicted vs. observed age of death for eight separate machine learning algorithms used in our main analysis of the cohort of 1910.

E Additional Results

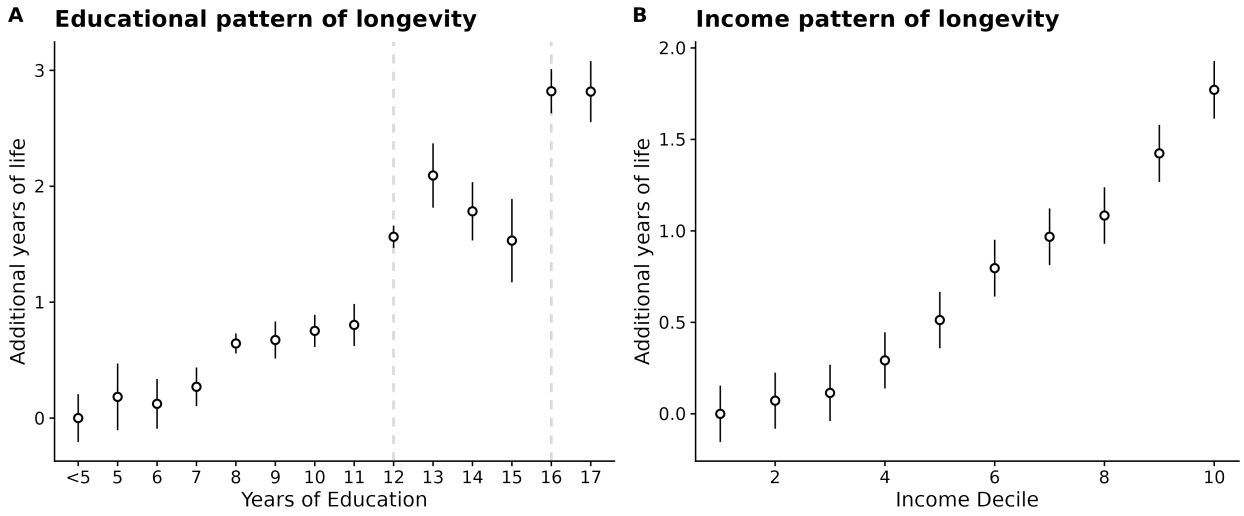


Figure A4: *(A)* Educational and *(B)* income gradients of longevity, calculated for individuals born in 1910. Error bars represent 95% uncertainty intervals.

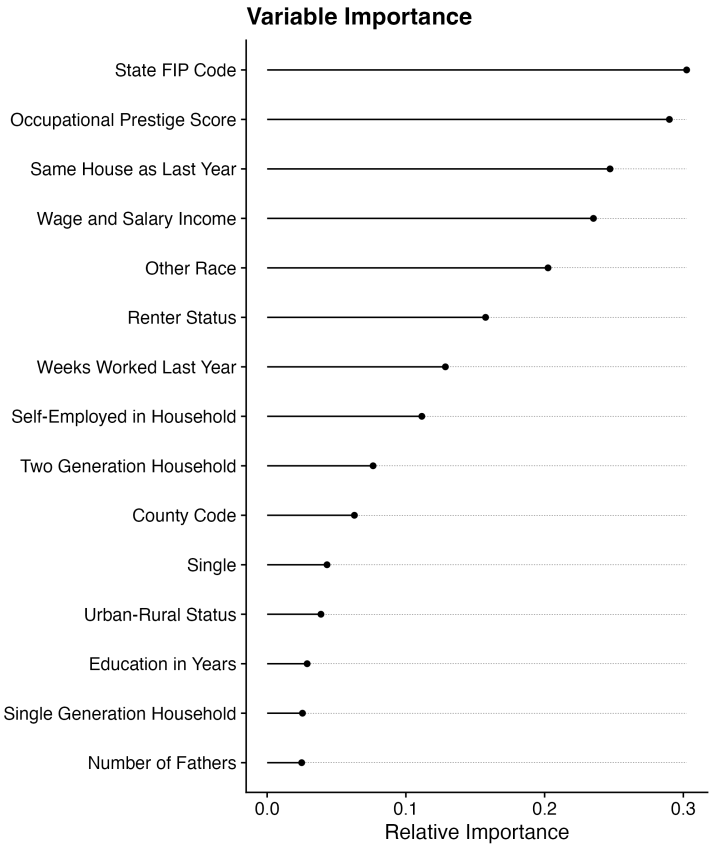


Figure A5: Variable importance for our main Superlearner model using a permutation approach. Specifically, variable importance is measured as the reduction in predictive accuracy (mean squared error) attributable to each variable if that variable was randomly permuted. This offers a different perspective, but it is sensitive to collinearity: highly correlated variables may mask each other's importance.

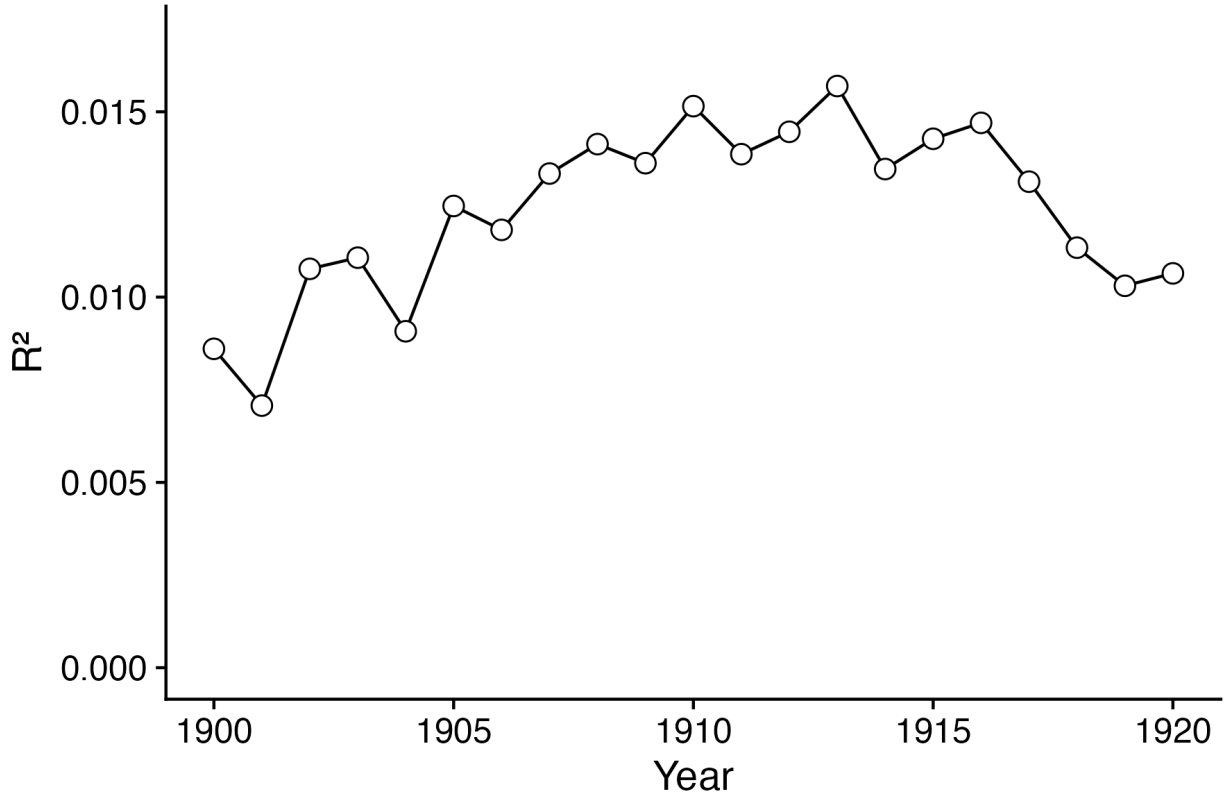


Figure A6: The estimated R^2 values from a replication of the main CenSoc-DMF prediction exercise, repeated separately for each birth cohort from 1900 to 1920. Performance is based on a gradient boosted model (GBM), our top-performing discrete machine-learning algorithm. We caution against overinterpreting trends, as each cohort will have: (1) covariates measured at a different point of the life course; (2) mortality observed for a different range of ages; and (3) cohort-specific exposures across the life course. In particular, directly testing whether covariates observed at earlier or later stages of the life course are more predictive is challenging because differences in predictive performance are also influenced by the different ages for which mortality is observed (e.g., the 1900 cohort has mortality observed from ages 75 to 105, while the cohort of 1920 has mortality observed from ages 55 to 85). Differences in predictive performance among cohorts likely reflect differences in age exposure rather than the intrinsic informativeness of later-life covariates. We include this figure to emphasize that predictability is low across all cohorts considered.

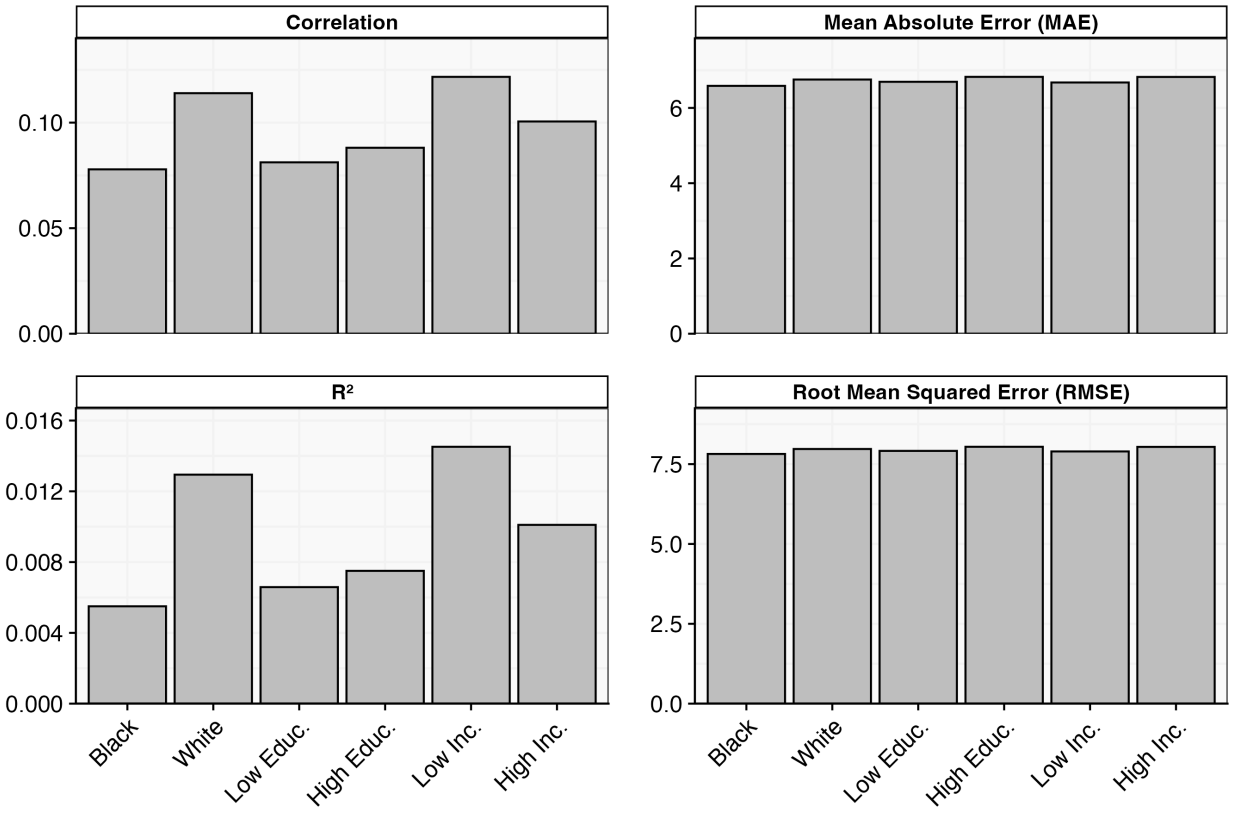


Figure A7: Performance metrics of the Superlearner algorithm by demographic and socioeconomic subgroups. Metrics include correlation, R^2, mean absolute error (MAE), and root mean squared error (RMSE); higher correlation and R^2 indicate better fit, while lower MAE and RMSE indicate lower prediction error. High and low education refer to educational attainment above or below sample median. High and low income refer to wage and salary income above or below sample median.

F REFORM Checklist

In order to promote transparency, we completed the REFORM checklist (Kapoor et al., 2024). The REFORM checklist, developed collaboratively by 19 researchers across the social, computer, and physical sciences, is a resource for promoting transparency and reproducibility in machine learning science.

F.1 Study Goals

1a. *State the population or distribution about which the scientific claim is made.*

In the primary analysis, this study makes claims about the predictability of lifespan using sociodemographic characteristics for the population of U.S. men born in 1910 dying between 1975 and 2005.

1b. *Describe the motivation for choosing this population or distribution.*

We focus on this population for several reasons. First, we are interested in mortality prediction in the United States, a country with well-documented group-level disparities in mortality. Second, we focus on these specific cohorts and age ranges due to data availability limitations—our mortality records only have complete coverage from 1975 to 2005. Importantly, our main analysis is restricted to men, as surname changes for women during marriage prevent accurate record linkage of women. Finally, we are interested in individuals who are adults when enumerated in the 1940 Census, as we want to measure covariates in early adulthood.

1c. *Describe the motivation for the use of ML methods in the study.*

The use of ML methods in this study is appropriate as our primary scientific interest

is assessing the predictability of individual-level mortality. Therefore, our interest is in maximizing the predictive power, rather than the explanatory power, of our models.

F.2 Computational Reproducibility

2a. *Describe the dataset used for training and evaluating the model and provide a link or DOI to uniquely identify the dataset.*

We use the publicly-available V3 CenSoc-DMF dataset (Goldstein et al., 2023).

2b. *Provide details about the code used to train and evaluate the model and produce the results reported in the paper along with link or DOI to uniquely identify the version of the code used.*

The models were fit using the SL3 package in R (Coyle et al., 2021). Code is available from a dedicated OSF repository [link].

2c. *Describe the computing infrastructure used.*

All computations were carried out on a 2023 MacBook Pro with an Apple M2 Pro chip, 16GB memory, and Sonoma 14.1 operating system.

2d. *Provide a README file which contains instructions for generating the results using the provided dataset and code.*

A README file is available as part of our replication package available on our OSF repository.

2e. *Provide a reproduction script to produce all results reported in the paper.*

A replication script is available on a dedicated OSF repository.

F.3 Data Quality

3a. *Describe source(s) of data, separately for the training and evaluation datasets, along with the time when the dataset(s) are collected, the source and process of ground-truth annotations, and other data documentation.*

The data are from the CenSoc project, which links the 1940 Census and Social Security mortality records (Breen, Osborne and Goldstein, 2023). Both our cohort and period datasets are split into a 75% training and 25% holdout dataset.

3b. *State the distribution or set from which the dataset is sampled (i.e., the sampling frame).*

The full sampling frame is all people enumerated in the 1940 Census who were successfully linked to death records between 1975 and 2005. For the cohort analysis, the data are first restricted to the birth cohort of 1910. For the period analysis, we take a random sample of 100,000 individuals between the ages of 54 and 95 in 1975. We take a random sample to minimize computational burden.

3c. *Justify why the dataset is useful for the modeling task at hand.*

This individual-level dataset contains many sociodemographic characteristics and age of death. In addition, the large size of this dataset and longitudinal structure allows us to investigate prediction from both a cohort and a period perspective.

F.4 Data Preprocessing

4a. *Describe whether any samples are excluded with a rationale for why they are excluded.*

For both cohort and period analyses, we exclude individuals with missing variables. We are interested in prediction for individuals with complete information on sociode-

mographic characteristics.

4b. *Describe how impossible or corrupt samples are dealt with.*

There are no impossible or corrupt samples.

4c. *Describe all transformations of the dataset from its raw form to the form used in the model, for instance, treatment of missing data and normalization—preferably through a flow chart.*

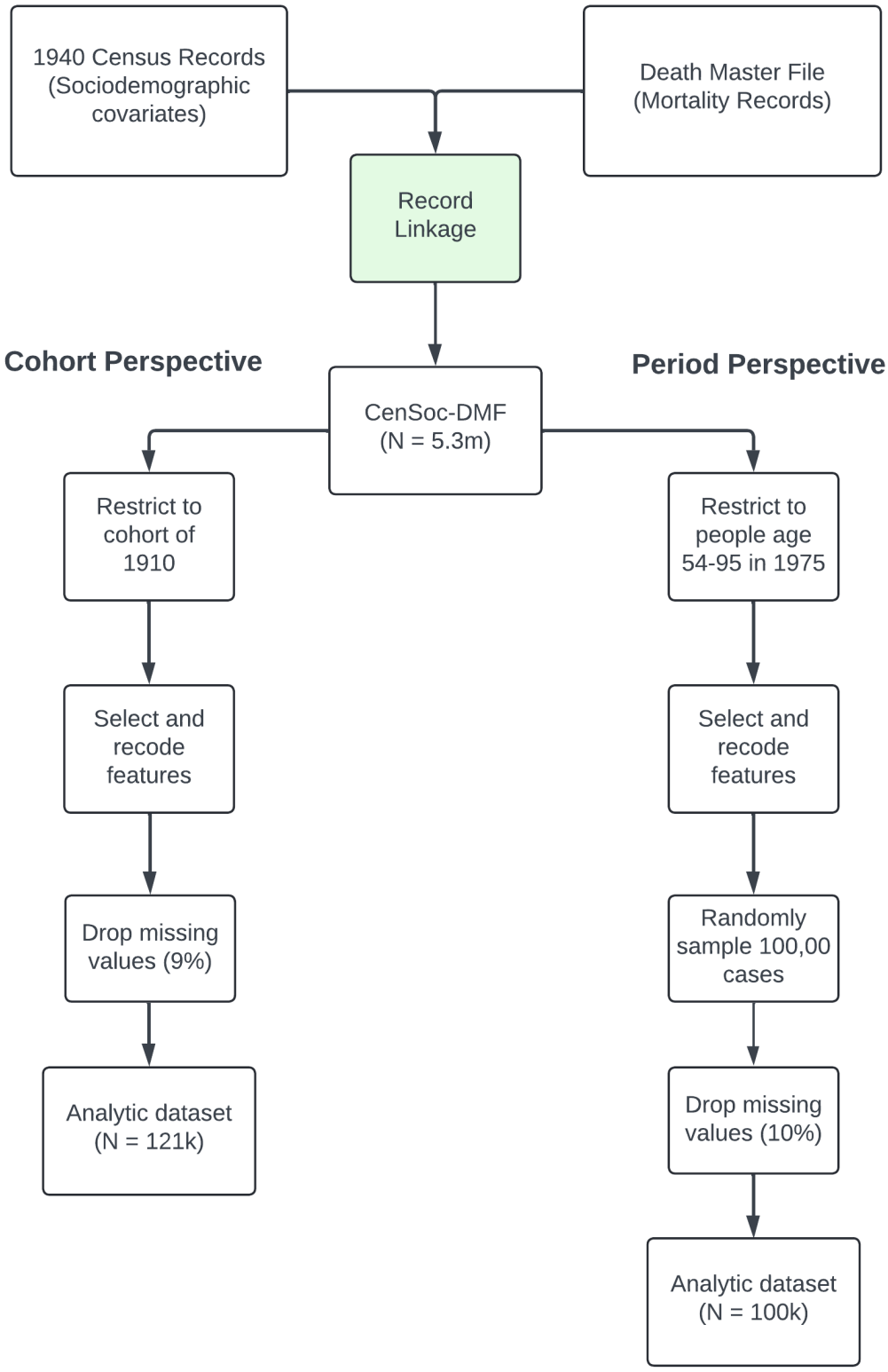


Figure A8: Data processing flowchart.

F.5 Modeling

5a. *Describe, in detail, all models trained.*

Our superlearner includes the following algorithms: gradient boosting machines, linear model, extreme gradient boosting, random forest regression, support vector machine, lasso regression, and multivariate adaptive regression splines (MARS). The model is fit using the SL3 package, an R package for fitting a Superlearner, a type of ensemble or "stacked" algorithm. The metalearner, which combines prediction from the set of discrete algorithms, is a non-negative ordinary least squares model.

5b. *Justify the choice of model types implemented.*

The key advantage of an ensemble learner with many different algorithms is that it reduces the need to rely on a given algorithm. Our candidate algorithms for our Superlearner were picked based on popularity and strong performance in similar prediction tasks.

5c. *Describe the method for evaluating the model(s) reported in the paper, including details of train-test splits or cross-validation folds.*

To assess the model performance, we randomly partition the data into a 75% training and 25% holdout training sets. Within the training dataset, the Superlearner algorithm uses 10-fold cross-validation to pick the best combination of algorithms.

5d. *Describe the method for selecting the model(s) reported in the paper.*

To select the models used in the Superlearner, we picked common machine learning algorithms that had performed well in other settings. We picked a diverse set of

algorithms to include in our ensemble per best practice (Phillips et al., 2023)

5e. *For the model(s) reported in the paper, specify details about the hyperparameter tuning.*

We conduct hyperparameter tuning to optimize the random forest and support vector machine algorithms. For random forest, we use a defined grid of hyperparameters, which includes the number of trees (100, 500, 1000), the number of variables to possibly split at each node (2, 3, 4), the minimum size of tree nodes (1, 5, 10), and the fraction of the data sampled for growing the trees (0.65, 0.8, 1). For Support Vector Machines (SVM), we use a structured search over a grid of two key hyperparameters: the cost parameter (0.1, 1, 10) and the gamma parameter (0.01, 0.1, 1). These parameters control the penalty of the error term and the influence of a single training example, respectively.

5f. *Justify that model comparisons are against appropriate baselines.*

Not applicable.

F.6 Data Leakage

6a. *Justify that pre-processing and modeling steps only use information from the training dataset (and not the test dataset).*

The pre-processing and modeling steps only use information available in the training dataset. The holdout data are exclusively used for prediction.

6b. *Describe methods used to address dependencies or duplicates between the training and test datasets.*

There are no dependencies or duplicates between the training and test (holdout) datasets.

6c. *Justify that each feature or input used in the model is legitimate for the task at hand and does not lead to leakage.*

No features or input in this model will lead to data leakage based on our study design. We randomly split our sample into a test and a holdout sample, and never use holdout data to train our models.

F.7 Metrics and Uncertainty

7a. *State all metrics used to assess and compare model performance. Justify that the metric used to select the final model is suitable for the task.*

For our cohort analysis predicting individual age of death, we used R^2 and mean absolute error (MAE). These metrics were selected as they are among the most popular metrics used in mortality prediction settings, allowing for the direct comparison to other studies. Further, these metrics are intuitive and straightforward to interpret.

7b. *State uncertainty estimates and give details of how these are calculated.*

We did not calculate estimates of uncertainty.

7c. *Justify the choice of statistical tests (if used) and a check for the assumptions of the statistical test.*

We did not perform any statistical tests.

F.8 Generalizability and Limitations

8a. *Describe evidence of external validity.*

Our findings align with past research on relative contribution of within-group differences and between-group differences to longevity (Caswell, 2009, 2023; Vaupel, 1988). However, assessing the predictability of mortality in other countries, ages, and time periods is an important direction for future research.

8b. *Describe contexts in which the authors do not expect the study's findings to hold.*

This study focuses on core sociodemographic characteristics within the context of a wealthy, developed country. We expect the broad patterns observed here to generalize across similar high-income settings with established data infrastructure systems and high social stratification. However, the extent to which these results would hold in low- and middle-income countries, where mortality patterns, data quality, and the salience of social determinants may differ, remains an open question.