ROCKWOOL Foundation Berlin
Institute for the Economy and the Future of Work (RFBerlin)

DISCUSSION PAPER SERIES 234/26

Socioeconomic Inequality in Longevity: A Multidimensional Approach

Paul Bingley, Claus Thustrup Kreiner, Benjamin Ly Serena

www.rfberlin.com SEPTEMBER 2026

Socioeconomic Inequality in Longevity: A Multidimensional Approach

Authors

Paul Bingley, Claus Thustrup Kreiner, Benjamin Ly Serena

Reference

JEL Codes: I14

Keywords: Life Expectancy, Inequality, Machine Learning

Recommended Citation: Paul Bingley, Claus Thustrup Kreiner, Benjamin Ly Serena (2026): Socioeconomic Inequality in Longevity: A Multidimensional Approach. RFBerlin Discussion Paper No. 234/26

Access

Disclaimer

Imprint

RFBerlin
ROCKWOOL Foundation Berlin –
Institute for the Economy
and the Future of Work

Gormannstrasse 22, 10119 Berlin
Tel: +49 (0) 151 143 444 67
E-mail: info@rfberlin.com
Web: www.rfberlin.com

Socioeconomic Inequality in Longevity: A Multidimensional Approach*

Paul Bingley[a]

Claus Thustrup Kreiner[b]

Benjamin Ly Serena[c]

Revised version: August 2026

Abstract

Socioeconomic inequality in longevity is typically measured using a single indicator, such as education or income. When indicators are imperfect proxies for socioeconomic status, this approach understates inequality. Using multiple indicators can improve measurement by better capturing latent socioeconomic status. We combine education, income, occupation, wealth, and IQ scores using machine learning. Using Danish population-wide data spanning 40 years, we estimate life expectancy at age 40 for the 1942–44 birth cohorts. Individuals at the top of the socioeconomic distribution live nearly 25 years longer than those at the bottom. The estimated gradient is 50–150% steeper with multiple indicators.

Keywords: Life Expectancy, Inequality, Machine Learning

JEL Classification: I14

---

* We are grateful to James Banks, Thomas Høgholm Jørgensen, participants at the Nordic Public Policy Symposium, and three anonymous referees for valuable feedback and to Erik Simonsen and Thorkild I.A. Sørensen for assistance in obtaining access to military recruitment data on IQ test scores. We thank Frederikke Veggerby Andersen for competent research assistance. We acknowledge the financial support of the Danish National Research Foundation (grant DNRF134) and the Novo Nordisk Foundation (grant NNF17OC0026542).

[a] The Danish Center for Social Science Research, 1052 Copenhagen K, Denmark. E-mail: pab@vive.dk.

[b] Center for Economic Behavior and Inequality, Department of Economics, University of Copenhagen, 1353 Copenhagen K, Denmark. E-mail: ctk@econ.ku.dk.

[c] The ROCKWOOL Foundation Research Unit, 1472, Copenhagen K, Denmark. E-mail: blse@rff.dk.

I. Introduction

Individuals with high socioeconomic status (SES) tend to live longer than those with low SES, and this longevity gap has been widening in many countries.[2] Together with rising economic inequality and declining social mobility (Piketty & Saez 2014; Chetty et al. 2017), this development has received increasing attention among policymakers (OECD 2018, 2022; UN 2022, 2023).

Empirical studies of the association between individuals' life expectancy and their SES typically rely on a single SES indicator, such as years of education or position in the income distribution.[3] These studies consistently find a significant SES gradient. Still, this inequality is underestimated unless the indicator is a perfect measure of latent SES.

In this paper, we combine multiple SES indicators using machine learning to capture as much of the SES gradient in life expectancy as possible.

We start by showing conceptually that the use of multiple SES indicators yields the best possible measure of SES inequality in longevity, provided that two conditions hold: each indicator adds unique information about latent SES (*relevance condition*) and does not predict longevity through channels unrelated to SES (*validity condition*). Using multiple indicators recovers more SES-related variation, bringing measured inequality closer to the true inequality associated with latent SES.

We use machine learning to estimate the SES gradient because it allows for nonlinearities, discontinuities, and interactions between the indicators in a data-driven way while avoiding

[2] Deaton 2013; NAS 2015; Case & Deaton 2015, 2020; Currie & Schwandt 2016; Chetty et al. 2016; Auerbach et al. 2017; Bor et al. 2017; Kreiner et al. 2018; Mackenbach et al. 2018; Couillard et al. 2021; Dahl et al. 2024; Danesh et al. 2024; Hagen et al. 2025.
[3] Meara et al. 2008; Mackenbach et al. 2008, 2018; Chetty et al. 2016; Kreiner et al. 2018; OECD 2018; Hederos et al. 2018; Kinge et al. 2019; Case & Deaton 2021; Milligan & Schirle 2021; Dahl et al. 2024; Milligan 2024.

overfitting. Hence, it allows us to exploit rich, multi-dimensional SES data while distinguishing meaningful patterns from noise in the data.

To conduct the empirical analysis, we merged Danish administrative records on key indicators commonly used to measure SES (Brown et al. 2019), including education, income, occupation, and wealth. For men, we also observe IQ scores from tests as part of compulsory military service assessments. To track mortality over the longest possible age span, we follow all individuals in birth cohorts 1942-44 from age 40 to 78 by their baseline SES. Based on this, we estimate differences in cohort life expectancy at age 40.

When combining all SES indicators, we find a gap in life expectancy of almost 25 years between individuals at the opposite extremes of the SES distribution for both men and women. Across the distribution, moving up ten percentiles is associated with an average increase in life expectancy of 1.6–1.7 years, and SES accounts for 12% of the total variance in life expectancy for men and 8% for women.

These findings contribute to the literature (op. cit.) measuring socioeconomic disparities in life expectancy. While such disparities are inherently underestimated when SES is measured imperfectly, our approach of combining all the available SES information satisfying the relevance and validity conditions approximates the true extent of SES inequality as closely as possible. We find that the SES gradient becomes 50–150% steeper relative to single-factor approaches using income or education alone, as is commonly done. Similarly, the share of variance in life expectancy explained by SES increases by 100–500%. We show that these substantially higher estimates persist across a wide range of robustness and validation checks and that our approach performs better than alternative multi-factor approaches.

Many studies may not be able to combine as much SES information as we do here. Therefore, we also analyze SES inequality when only a subset of indicators is available and investigate the extent to which indicators can substitute for each other. Across combinations of indicators, going from one to two indicators substantially increases the share of variance explained, but it remains well below that explained when all indicators are included. We find that wealth provides information that is difficult to substitute with the other indicators. Relatedly, wealth is the best single SES measure, accounting for 8% of the variance in life expectancy for men and 5% for women. This finding, based on population-wide administrative records, corroborates recent evidence from smaller samples based on surveys, pointing to larger inequalities in life expectancy by self-reported wealth than by income or education (Glei et al. 2022).

Our analysis relates to other studies that use machine learning to predict mortality (Ganna & Ingelsson 2015; Einav et al. 2018; Puterman et al. 2020). Our analysis is distinct in that we do not focus on obtaining the best possible mortality prediction. This would involve including as much information as possible, for example, prior health information (Einav et al. 2018). Rather, we use machine learning to provide a new measure of the socioeconomic gradient by including SES-relevant information and excluding variables that may capture mortality differences unrelated to SES (violating the validity condition).

The following sections describe the conceptual framework, data, estimation method, empirical results, and robustness analyses, before concluding in Section VII.

II. Measurement with Multiple Socioeconomic Indicators

Our approach is to estimate

$$Y = f(\mathbf{x}) + u, \tag{1}$$

where Y is the outcome, $\mathbf{x}$ is a vector of observed socioeconomic indicators, $f(\mathbf{x})$ is the prediction of Y based on these indicators, and u is residual variation in Y. We estimate $f(\mathbf{x})$ using flexible machine learning methods, which allow for nonlinearities, discontinuities, and interactions between indicators.

If socioeconomic inequality in longevity is defined as the share of variation in longevity that can be predicted by observed SES indicators, eq. (1) directly delivers the relevant object: $\text{Var}\big(f(\mathbf{x})\big)/\text{Var}(Y)$. This interpretation would treat the observed socioeconomic indicators themselves as defining the SES-related component of longevity.

More naturally, we consider each SES indicator as an imperfect proxy for an underlying latent SES. Let S be a uniform latent rank variable reflecting the true SES position of individuals, and suppose that the outcome is related to S according to

$$Y = h(S) + \nu, \tag{2}$$

where $h(S)$ is a monotonic function reflecting the true association between SES and the outcome, while ν is outcome variation unrelated to SES. In this setting 'true' SES inequality is defined as $\text{Var}\big(h(S)\big)/\text{Var}(Y)$. We want to know how well this is captured by our estimate $\text{Var}\big(f(\mathbf{x})\big)/\text{Var}(Y)$ when each observed SES indicator is an imperfect proxy for S:

$$x_k = g_k(S) + \epsilon_k. \tag{3}$$

Here, $g_k(\cdot)$ denotes the SES-related component of indicator x_k, while ϵ_k captures variation in x_k unrelated to latent SES.

Under the assumption that the SES indicators do not provide any information on the outcome beyond its effect through SES, i.e.

$$\mathbb{E}[\nu \mid \mathbf{x}] = 0, \tag{4}$$

we obtain

$$\mathrm{Var}\big(f(\mathbf{x})\big) = \mathrm{Var}(\mathbb{E}[\mathrm{h(S)} \mid \mathbf{x}]) \leq \mathrm{Var}\big(\mathrm{h(S)}\big). \tag{5}$$

The inequality follows from the law of total variance implying $\mathrm{Var}\big(h(S)\big) = \mathrm{Var}(\mathbb{E}[h(S) \mid \mathbf{x}]) + \mathbb{E}[\mathrm{Var}(h(S) \mid \mathbf{x})]$, where the first term is the SES-related variation in the outcome that is recoverable from the observed indicators (between-**x** variation), while the second term is residual latent SES-related variation (within-**x** variation).

Equation (5) shows that SES proxies will tend to understate the latent SES inequality. Using multiple SES indicators can reduce this attenuation bias if each additional indicator provides additional information about SES conditional on the existing indicators (*relevance condition*). At the same time, the added indicators should not predict longevity through channels unrelated to SES, as stated in equation (4) (*validity condition*). These relevance and validity requirements apply equally to single-factor and multi-factor approaches, but the single-factor approach is more vulnerable to attenuation.

This reasoning motivates our empirical strategy of combining several relevant SES indicators to recover a larger share of the latent SES-related variation in longevity. At the same time, we deliberately exclude variables such as prior health incidents, even though they are highly predictive of longevity, because they may capture outcome-relevant variation unrelated to socioeconomic differences among individuals. We further elaborate on the fulfillment of the validity condition in Section VI, which also describes results from using alternative multi-factor approaches.

III. Data and Sample Selection

We use population-wide Danish administrative data on SES and mortality during 1980-2022. Using personal identifiers assigned at birth (CPR number), we link population-wide administrative registers on gender, age, income, education, wealth, occupation, and mortality. For men, we also link the administrative data with military records containing the results of mandatory conscription IQ tests.

We base our empirical analysis on cohort life expectancy where we follow mortality of individuals born between 1942 and 1944 (aged 40 in the years 1982-1984).[4] We use this sample because we want the longest possible panel of mortality records—from age 40 to 78 years—to obtain precise life expectancy predictions. We measure SES two years before we start analyzing mortality to reduce the potential for reverse causality (Chetty et al. 2016; Kreiner et al. 2018). Hence, we measure all socioeconomic indicators at age 38 and keep these measures fixed throughout the analysis. Thus, mortality prediction at every subsequent age—from 40 to 78—is based on the same baseline SES information measured at age 38.

[4] We use cohort life expectancy to avoid introducing measurement issues that could interfere with the conceptual measurement exercise and the comparison of the multi-factor approach with the single-factor approach. Period life expectancy is more susceptible to bias because of reverse causality, socioeconomic mobility, and changes in group composition across ages. For example, when based on education, period life expectancy can be biased by the large expansion of education over time (Hendi 2017) and when based on income, the gradient is upwardly biased because of income mobility (Kreiner et al. 2018).

The selected cohorts include 118,026 men and 113,221 women. We further restrict the sample to individuals with non-missing education, income, wealth, and occupation (and for men, IQ) at age 38 and who live in Denmark throughout the analysis period (to age 78 or death), leaving 89,927 men and 101,466 women.[5,6] The larger share of discarded observations among men is due to incomplete coverage in the military records. In the following, we describe how we measure each socioeconomic indicator used in the analysis (see the appendix for additional details). We focus on individual rather than household measures of SES as IQ scores are only available for men. Results using household variables are provided in the appendix.

Income: We measure income net of universal transfers in line with previous work (e.g., Kreiner et al. 2018; Dahl et al. 2024). Hence, we exclude cash assistance and disability pensions, but include unemployment benefits, which are tied to previous earnings.[7]

Education: We measure education as the length of education in years and dummies for individual educational qualifications (e.g., engineer, economist, etc.).

Wealth: We measure wealth as the difference between assets and liabilities. Assets include third-party reported housing assets, bank account balances, stocks and bonds, equity in private companies, and self-reported values of cars, boats, and campers.[8] Liabilities include bank loans, mortgages, student debt, and debt to the tax authorities. Housing assets are known to be undervalued by tax authorities, while housing liabilities (mortgages, etc.) reflect the actual price of buying the house. To address this issue, we follow previous research (Boserup et al. 2016; Leth-Petersen 2010) and scale up the tax-assessed value of housing assets by 20% to match average

[5] The occupation classification used includes categories for unemployment and retirement. Hence, the analysis sample is not restricted to individuals in employment at age 38.
[6] We exclude the latter group because deaths occurring outside Denmark are unobserved. In the appendix, we show that the results are similar if we include all individuals, including those with missing SES information.
[7] This income measure is more closely linked to earnings capacity than total income and is therefore less likely to rank individuals with temporarily low incomes below those receiving permanent disability pensions.
[8] The self-reported components of recorded assets are negligible (see e.g., Appendix Figure A1 in Boserup et al. 2017).

actual house prices. During our analysis period, the combined wealth of married couples is recorded in the husband's name. Therefore, we compute the total wealth of married couples and divide it equally between the two partners.

IQ: At age 18, all Danish men are requested to attend a regional conscription board to assess their fitness for military service. The Danish Conscription Database contains information on these assessments (Christensen et al. 2015). Our IQ measure is the Danish Armed Forces Qualification Test, which was developed for military recruitment to measure cognitive ability. The correlation coefficient between these scores and the standard Wechsler Adult Intelligence Scale is 0.82 (Mortensen et al. 1989). Between five and ten percent of men were deemed unfit for service and exempt from assessment because of medical conditions, e.g., intellectual disability or epilepsy (Green 1996). Therefore, around 15% of men in the selected cohorts do not have a recorded IQ measure.

Occupation: We measure occupation by including dummy variables for occupation codes, using a Danish classification comparable to the Disco coding system. Additionally, we include dummies for broader occupational categories based on this classification, for example self-employed with fewer than 20 employees.

In total, we include 798 different variables (most of which are dummy variables for education and occupation codes).

IV. Estimation Approach

We are interested in capturing all relevant SES-related variation in life expectancy without imposing a predefined SES grouping. Therefore, we let the data tell us which interactions and combinations of SES indicators are most predictive for life expectancy. If longevity was directly observed in the data, we could achieve this goal by using a flexible and data-driven method, such as machine learning, to predict longevity using all the available SES indicators. However, while our data covers mortality over four decades from age 40 to 78, the age of death of those still alive by 78 is unobserved. To address this issue, we proceed in two steps. In the first step, we use machine learning to predict mortality in the ages we observe. In the second step, we use survival models to extrapolate these mortality estimates beyond the observed ages and estimate cohort life expectancy.

Step 1: Predicting mortality at observed ages

We start by randomly allocating the individuals in our data to a training sample (60%) and a test sample (40%). We use the training sample to train the model, while the test sample is used to validate the estimates.

We construct panel datasets in which we follow individuals from age 40 to 78 or until death. From these, we create balanced samples with 50% mortality by randomly oversampling person-age observations in which a death occurs. We do this because fitting an algorithm is easier when the two possible outcomes—dead and alive—are equally likely (Einav et al. 2018).

We use an eXtreme Gradient Boosting (XGBoost) algorithm to predict mortality in the balanced training dataset. Like random forest, this algorithm works by creating and averaging predictions across many decision trees.[9] In addition to our SES indicators, measured at age 38, we include age in the model to allow predicted mortality to vary by age.

[9] Unlike the random forest algorithm, where differences in trees come from different random subsets of the data, XGBoost creates additional trees by improving upon (boosting) previous trees. In this way, the model becomes more accurate with every iteration. The algorithm is widely used for a broad range of applications and has received acclaim for its performance (Bentéjac et al. 2021). XGBoost relies on several hyperparameters: the learning rate, maximum tree depth, regularization parameters, and the number of trees. These parameters control the trade-off between exploiting all available information and the risk of overfitting. To find the optimal parameter values, we perform hyperparameter tuning using five-fold cross-validation and Bayesian Optimization (Frazier 2018). The various steps, including tuning the hyperparameters and running the XGBoost algorithm, use random draws to split the data and to sample candidate parameter values. Hence, the results vary slightly each time the code is executed. To smooth this variation, we run the previous steps 50 times, using different seeds for the random draws, and then take the average across iterations.

Because we run the model on a balanced sample with 50% mortality, the predicted mortality rates are artificially high. We therefore rescale the predictions. Specifically, in the original training sample, we fit a logistic regression of mortality on a fifth-order polynomial of the machine-learning predictions, allowing this relationship to vary with age (Einav et al. 2018). The fitted mortality rates then match average mortality in the population.

Step 2: From predicted mortality profiles to life expectancy

We now have individual-level mortality predictions from age 40 to 78. To extrapolate these mortality predictions and obtain life expectancy predictions, we first collapse these 39 parameters into two variables that can be included in a survival model. Specifically, we run standard individual-specific OLS regressions of log predicted mortality on age. We then use the estimated intercepts and slopes from these regressions as two explanatory variables in a Gompertz survival model (Gompertz 1825).

The survival model assumes that the hazard rate λ evolves exponentially with age:

$$\log(\lambda) = (\alpha_0 + \alpha_1 \widehat{w}_i) + (\beta_0 + \beta_1 \hat{z}_i) \cdot age,$$

where $\widehat{w}_i$ and $\hat{z}_i$ are the OLS estimates of the intercept and slope of predicted log mortality.[10]

We estimate the survival model using the training sample and use the estimated coefficients to predict mortality, survival, and life expectancy in the test sample. Because the test sample is not used in the estimation, it allows us to assess the out-of-sample accuracy of our final models (see Appendix Figure 1).[11]

[10] Since these individual-level intercepts and slopes are based on Gompertz regressions of predicted log mortality, the estimated α_1 and β_1 parameters are close to unity.

[11] Note that our approach allows covariates to change the slope of log mortality, in contrast to the proportional hazards model. More direct implementations of survival analysis in machine learning, such as DeepSurv (Katzman et al. 2018) and standard survival implementations in XGBoost rely on the assumption of constant slopes, and are, therefore, more parametrically restrictive. The added flexibility of our approach is crucial when studying SES differences in mortality, as the intercept and slope of log mortality are typically inversely related across subgroups, known as the compensation effect of mortality (Gavrilov & Gavrilova 1991).

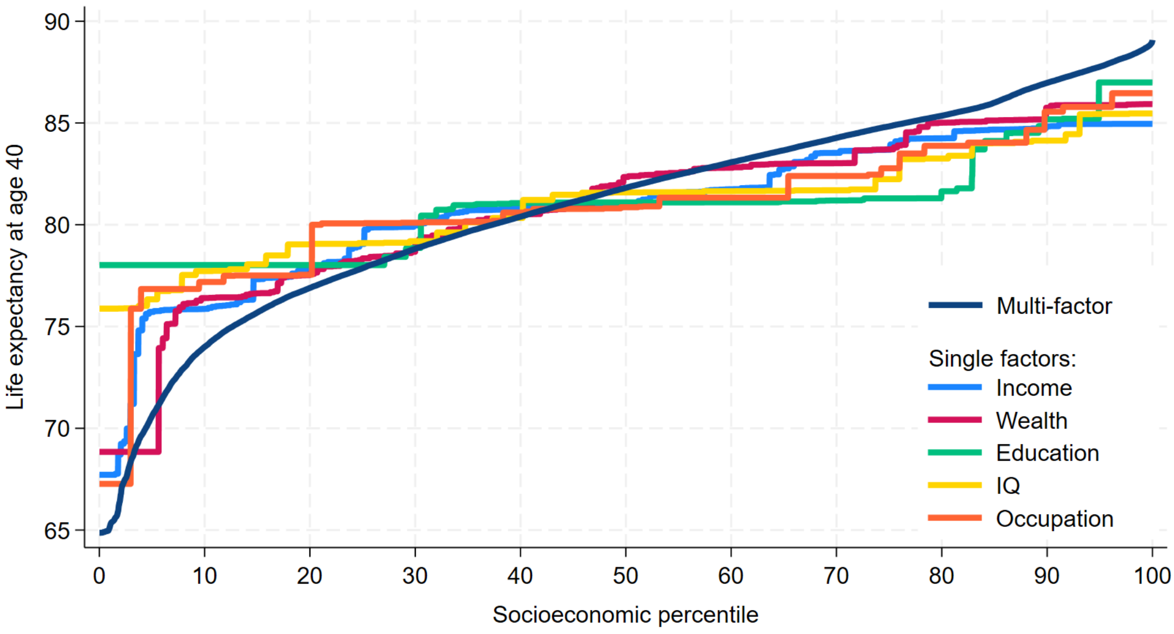
V. Empirical Results

Figure 1 shows cohort life expectancy at age 40 by socioeconomic rank for men and women. The figure is constructed in the following way: For the line denoted Multi-factor, we first use all socioeconomic information to run the machine-learning approach described above. Using the estimated algorithm, we predict life expectancy for individuals (in the test sample) and rank people on a scale from 1 to 100 (percentile ranks) according to the prediction. Finally, we plot the predicted life expectancy of the individuals by their rank as shown in Figure 1. By construction, the graph is increasing, but it can have a stepwise profile because the machine-learning algorithm divides people into groups with distinct differences in life expectancy based on their SES indicators.

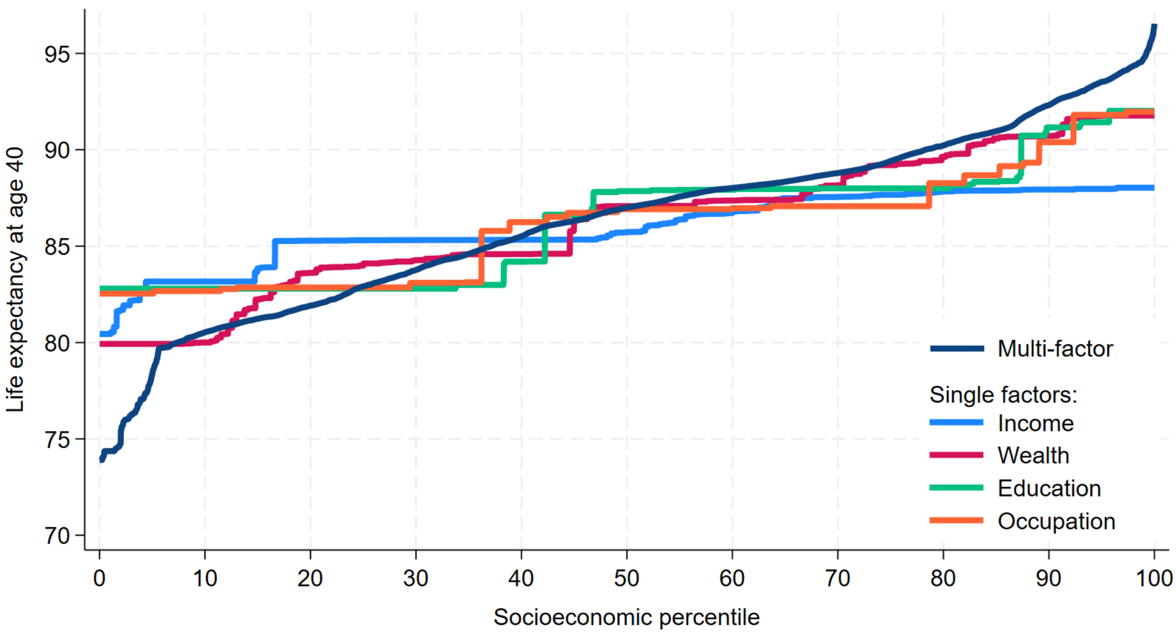
When using a single socioeconomic indicator (income, wealth, education, occupation, or IQ) to measure the socioeconomic gradient, we apply the same machine-learning approach. This ensures that the results of our multi-factor approach (using multiple SES indicators) and the standard single-factor approach are compared on equal terms. Machine learning can also improve the single-factor approach when the relationship between an SES indicator and life expectancy is non-monotonic. Individuals with the lowest observed incomes or wealth may, for example, be better off than those just above them (Brewer et al. 2017). Previous studies commonly address this issue by excluding observations at the bottom of the distribution (Chetty et al. 2016; Kinge et al. 2019; Hagen et al. 2025). In our setting, the issue is most relevant for wealth (see Appendix Figure 2), where individuals with wealth close to zero have lower life expectancy than those with negative wealth, a group that includes self-employed individuals with large business losses. The machine-learning algorithm accommodates this non-monotonicity by assigning individuals with negative wealth higher SES ranks than those with near-zero wealth, allowing us to retain both groups in the analysis.

Figure 1. Life expectancy at age 40 by SES

(a) Men



(b) Women



Note: The figure plots cohort life expectancy at age 40 by single-factor and multi-factor measures of socioeconomic status. The socioeconomic percentile (on the x-axis) is calculated as the position in the distribution of predicted life expectancy based on the selected SES measure. Multi-factor refers to SES measured using the most predictive combination of all the listed single factors for life expectancy.

When including all indicators, life expectancy at age 40 for men starts at 65 years for those with the lowest SES, gradually increasing to 89 for those with the highest SES (see Figure 1 and Table 1). This difference in life span of 24 years is much greater than the difference obtained when using only one of the indicators. For the most used indicators of SES in this context, education and income, the life span difference is 9 and 17 years, respectively. The results for IQ, wealth, and occupation are similar: 9, 17, and 19 years. We reach the same conclusions when comparing lifespan differences between the 10th and 90th percentiles of socioeconomic rank. The multi-factor approach yields a 13-year difference, while the single-factor approach yields differences from 6 to 10 years.

Interestingly, the life-expectancy estimates based on education alone come closest to the multi-factor estimate at the top of the distribution but diverge substantially at the bottom, whereas the opposite pattern holds for income. More generally, indicators of economic resources, particularly income and wealth, appear to capture disadvantage at the bottom of the predicted-longevity distribution better than indicators of human capital, such as education and IQ, consistent with findings in Hagen et al. (2025).

Table 1 also reports the estimated SES gradient in life expectancy. For example, moving up 10 percentiles in the income-life expectancy distribution is associated with an increase in life expectancy of 1.16 years, while the same move in the multi-factor distribution corresponds to a gain of 1.72 years. Hence, the SES gradient increases by 50% when using the multi-factor approach. When comparing to using education alone, the increase is even larger at 120%.

Finally, we measure SES inequality by computing the ratio of the variance in predicted life expectancy across SES groups to the population variance in age at death. We find that about 3% of the variance in life expectancy is accounted for by education. This percentage increases fourfold to 12% when using all socioeconomic indicators. For the single-factor approach, income and wealth give very similar results when comparing Min-Max and P10-P90 differences, but wealth accounts for more of the overall variance with an estimate close to 8%. This reflects the fact that the wealth-based estimates track the multi-factor estimates more closely throughout the distribution.

Table 1. Life expectancy differences and explained variance by SES indicators

	Multi-factor	Education	Income	Wealth	IQ	Occupation
Men						
Min-Max	65 − 89	78 − 87	68 − 85	69 − 86	76 − 85	67 − 86
P10-P90	74 − 87	78 − 85	76 − 85	76 − 86	78 − 84	77 − 86
Linear gradient	0.172	0.078	0.116	0.136	0.084	0.107
SES share (%) of var.	11.8	2.7	5.8	7.9	2.7	5.4
Women						
Min-Max	74 − 97	83 − 92	80 − 88	80 − 92		83 − 92
P10-P90	81 − 92	83 − 91	83 − 88	80 − 91		83 − 90
Linear gradient	0.156	0.100	0.058	0.121		0.093
SES share (%) of var.	8.4	3.7	1.3	5.0		3.2

Note: The table reports inequality in cohort life expectancy at age 40 by single-factor and multi-factor measures of socioeconomic status. The upper pane reports results for men, the lower results for women. Within each pane, the first row shows the minimum and maximum predicted life expectancy in the sample when based on the selected SES measure (reported in columns). The second row reports the 10th and 90th percentiles. The third row reports the slope of a linear regression line across percentiles of the SES measure. The fourth row reports the share of the overall variation in life expectancy accounted for by the selected SES measure. This is calculated as the ratio of the between-SES-group variance in life expectancy to the overall variance in age at death.

The difference between the multi-factor and single-factor estimates is even greater for women (despite the availability of only four SES indicators). Women can expect to live longer than men (note the difference in scale on the y-axis in Figure 1). Still, the difference in life expectancies between low-SES and high-SES women is 23 years when including all socioeconomic indicators, almost the same as for men. In contrast, the difference between low-SES and high-SES women is only around 9 years when using education, income, or occupation alone and only 12 years when using wealth. Hence, the ability of the multi-factor approach to combine the SES information increases the predicted gap in life expectancy of women significantly compared to using any of the SES indicators alone.

Table 1 shows that income and occupation account for less of the variance in life expectancy for women than men, possibly due to less labor market participation and less full-time work for these cohorts of women. For education, the share of the variance in life expectancy attributed to SES is higher for women than for men.

Summarizing across genders, the steepness of the SES gradient increases by 50-150%, and the explained variance in life expectancy increases by 100-500% when combining multiple indicators compared to using income or education alone.

From an external validity perspective, it is worth noting that, despite differences in the data and methods used across studies, our estimates based on single SES indicators are in the same ballpark as previous findings. For example, the P90-P10 gap in life expectancy by income is nine years for men and five years for women, which aligns with the results of Chetty et al. (2016). Similarly, we find that education accounts for 3% of the variance in life expectancy, which is consistent with van Raalte et al. (2012).

There may be several reasons why the multi-factor approach performs better. As noted above for the single-factor analysis, some indicators have greater explanatory power at the top of the distribution, while others have greater explanatory power at the bottom. The multi-factor approach takes these differences into account when making the best prediction.

We also find that having consistently high rankings across multiple socioeconomic indicators is much more predictive of high life expectancy than being at the top of any single indicator (see Appendix Table 1). Among men with a predicted life expectancy in the top 10%, around 40% also rank in the top 10% of the separate distributions of education, income, wealth, IQ, and occupation. In contrast, around 80% of them rank in the top 10% when individuals are ranked by their average position across all socioeconomic indicators. This pattern is even more pronounced at the very top of the distribution. Among the top 1% with the highest predicted life expectancy based on our multi-factor approach, only 5–15% are in the top 1% of any single socioeconomic indicator. In contrast, around 60% of them belong to the top 1% when ranked by the average of all indicators.

This also helps explain why the multi-factor gradient in Figure 1 does not flatten at the top like the single-factor gradients. For example, individuals with similarly high incomes may differ substantially in education, wealth, occupation, and IQ, yet receive the same predicted life expectancy in the income-based single-factor approach. The multi-factor approach distinguishes between individuals who rank highly only on income and those who rank highly across all SES indicators, thereby revealing a steeper longevity gradient.

We observe the same patterns for women and for individuals at the bottom of the distribution for both genders. Specifically, being in the bottom 10% ranked by the average of all indicators is more predictive of low life expectancy than being in the bottom 10% of any single indicator. A strength of the machine-learning approach is its ability to detect such patterns and exploit them when deriving the best prediction.

Combining a Limited Set of Socioeconomic Indicators

Our data allows us to study a comprehensive set of socioeconomic indicators. In Table 2, we examine whether a subset of indicators can approximate the full multi-factor estimate and whether individual indicators can substitute for each other in the prediction of longevity.

The upper pane of Table 2 shows the share of the variance attributed to SES when we exclude one of the SES indicators from the multi-factor analysis. For men, the share is almost unchanged when removing education, IQ, or occupation, reflecting that each of these indicators add little additional SES information conditional on the other four indicators. For instance, while occupation alone accounts for 5.4% of the variation in life expectancy (according to Table 1), excluding it from the multi-factor analysis decreases the share from 11.8% to 11.4%.

Income is somewhat more important, but the largest decrease occurs when we remove wealth. In this case, the share accounted for by SES drops to 8.4%, demonstrating that the information contained in wealth cannot be substituted by the four other indicators.

Table 2. Explained variance in life expectancy by SES using limited sets of indicators

SES share (%) of variance	Education	Income	Wealth	IQ	Occupation
1. Excl. one indicator					
Men	11.4	10.8	8.4	11.2	11.4
Women	7.1	8.0	6.0		8.1
2. Combining two indicators					
Men					
Education		6.6	9.1	3.2	6.0
Income			10.2	6.8	7.6
Wealth				9.2	9.7
IQ					6.1
Women					
Education		5.0	7.1		5.3
Income			5.9		3.9
Wealth					6.8

Note: The table reports the share of the overall variation in cohort life expectancy at age 40 accounted for by different measures of SES, based on combinations of SES indicators. The estimates are calculated as the ratio of the between-SES-group variance in life expectancy to the overall variance in age at death. The upper pane reports estimates based on the most predictive combination of all SES indicators (multi-factor approach), excluding the indicator listed in the column title. The lower pane reports estimates based on the combination of only two SES indicators, given by row and column titles.

For women, the drop in the share when excluding income or occupation is small, aligning with our previous findings that these indicators are less informative for women than for men. However, like men, the share drops most when excluding wealth.

The lower panel of Table 2 shows the share of the variance in life expectancy attributed to SES when combining pairs of SES indicators. The results show that moving from the single-factor estimates in Table 1 to just two indicators substantially increases the share of variance attributed to SES. For example, we saw previously that wealth is the most predictive single indicator, accounting for 7.9% of the variance. This percentage increases to around 10% when we add income or occupation, illustrating that with just two indicators, we can approach the share accounted for when using all five indicators. For women, the most predictive two-factor combination is wealth and education, which accounts for 7.1% of the variance compared to 8.4% when including all indicators.

VI. Robustness

Validity condition: A potential concern is that our multi-factor approach, which uses outcome predictions to construct SES ranks, may capture variation in life expectancy unrelated to SES, thereby violating the validity condition. For example, if mortality is high among a high-income group with a risky occupation, these individuals may be assigned a low "predicted rank" despite having relatively high SES. Our finding that the predicted longevity rank is strongly and monotonically related to the underlying SES indicators (Appendix Figure 2) alleviates concerns that the algorithm is primarily sorting individuals according to idiosyncratic mortality risks unrelated to SES.

To provide further evidence, we repeat the ranking exercise using the income of the children of individuals in our sample as an alternative outcome, while retaining the same parental SES indicators. Because children's income is not a mortality outcome, this exercise provides an independent validation of the socioeconomic content of the ranking. The resulting ranks are strikingly similar to those obtained using longevity as the outcome and predict longevity gradients similar to those obtained with the original longevity-based ranking (see Appendix Figure 3). We interpret these findings as evidence that the algorithm primarily extracts socioeconomic information rather than outcome-specific variation.

Alternative multi-factor approaches: An alternative to our multi-factor approach is to first estimate latent SES, S, without using outcome information, for example through common factor analysis or principal component analysis, and then estimate the association between the outcome and the estimated S. These approaches exploit the variation shared across SES indicators.

Compared with our prediction-based approach, this may reduce bias from indicators that predict longevity for reasons unrelated to SES, but it also discards SES-relevant variation that is unique to a particular indicator. We examine this alternative in Appendix Figure 4 and show that it yields estimates similar to, and sometimes smaller than, those obtained using a single SES indicator. Hence, these methods fail to capture the full extent of inequality in longevity accounted for by the SES indicators.

Our prediction-based approach combines the SES indicators in the way that best predicts mortality. We use machine-learning methods for this task because they are designed to maximize out-of-sample predictive performance in a flexible, data-driven, and non-parametric way while limiting overfitting. We show in Appendix Figure 5 that parametric regressions do not capture the full predictive potential of the SES indicators.

Sample selection: We exclude individuals who either have missing SES information at age 38 or leave Denmark during the analysis period. We conduct two complementary sensitivity analyses to assess the importance of these restrictions. In the first analysis, we reproduce our main results without imposing either restriction (see Appendix Table 2).[12] This will tend to bias predicted longevity upward because deaths of Danes living abroad are not observed. Despite this, the results are very similar to our baseline estimates, although predicted longevity at the top of the distribution is somewhat higher.[13]

In a second analysis, we focus on men excluded from the sample because they did not participate in the mandatory military recruitment assessments (15% of the sample) and therefore have missing IQ information. As this only applies to men, it may affect comparisons between men and women. Appendix Table 3 therefore repeats the analysis for men excluding IQ scores in

[12] XGBoost can accommodate missing explanatory variables directly without imputation. Intuitively, when a tree splits individuals according to a particular SES indicator, individuals for whom that indicator is missing are not dropped; instead, the algorithm places them on the side of the split that yields the best prediction.
[13] This likely reflects that this group is more likely to move in and out of Denmark, causing an upward bias.

the model and in the sample restrictions. The results are again very similar to the baseline estimates. Overall, these analyses suggest that our sample restrictions do not affect the results.

Variable definitions and measurement: We use individual-level SES indicators throughout because IQ is only available for men. In Appendix Table 4, we show results using household variables instead, e.g., the average income of spouses. For women, using household variables increases the share of variance accounted for by SES, showing that household-level SES indicators can be more predictive than individual-level SES indicators. For men, the results are broadly unchanged.

Following previous studies (Chetty et al. 2016; Kreiner et al. 2018), we measure SES two years before starting to analyze mortality to reduce the potential for reverse causality, which can occur if individuals are in poor health before we measure their SES. In Appendix Table 5, we remove individuals who receive disability insurance benefits at age 38 from the sample. This can reduce reverse causality but also remove some of the inequality. The share of variance accounted for by SES falls somewhat, but the relative difference between the multi-factor and single-factor approaches is very similar to Table 1.

Extrapolation: We observe individual mortality from age 40 to 78 and follow previous studies by estimating a Gompertz model of age-dependent mortality rates to derive life expectancy estimates. Appendix Figure 1 shows that the Gompertz model fits the data well. We obtain almost identical results if we instead use a Kannisto model (Gavrilov & Gavrilova 2019) to predict mortality rates (see Appendix Table 6).

We have also repeated our main analysis using survival until age 78 as the outcome, which we observe in the data. Appendix Figure 6 plots the observed survival rate against the machine-learning prediction of socioeconomic rank. It mirrors Figure 1 with the multi-factor approach giving a considerably steeper gradient. For example, among men, the multi-factor line shows that the survival probabilities range from 34% at the lowest SES level to 78% at the highest, compared to a 42%-74% span across income groups and a 57%-76% span across education groups.

VII. Conclusion

Socioeconomic inequality in life expectancy is typically measured as the association between life expectancy and a single socioeconomic indicator approximating latent SES. We ask whether the socioeconomic gradient becomes larger when combining several indicators. If the different SES indicators capture the same underlying association between SES and longevity, the multi-factor gradient need not exceed the gradient based on single indicators. Our analysis reveals that the socioeconomic gradient increases substantially when indicators are combined—by 50-150% relative to the single-factor gradients based on the most commonly used indicators, income and education. This suggests that standard estimates of SES differences in life expectancy fail to capture the full extent of inequality.

Information on multiple socioeconomic indicators in combination with mortality is often unavailable. In that case, our study can also provide useful information. Some examples are as follows: Education is good at capturing variation in the upper portion of the SES distribution but is less good at capturing it at the lower end. For income, it is the other way around. Overall, combining SES indicators improves predictive accuracy and reveals a steeper SES gradient. Wealth stands out as the most predictive single indicator, and the other indicators can only partially compensate for missing wealth information in the measurement of inequality in life expectancy.

Our results are useful for the measurement and monitoring of inequality but do not identify causal mechanisms or policies that would reduce longevity inequality. However, individuals with the lowest predicted longevity are particularly disadvantaged in terms of income and wealth, complementing recent findings by Hagen et al. (2025). Policies targeting those with limited economic resources may therefore also reach those with the lowest predicted longevity. Although we cannot infer that such policies would increase life expectancy, they would compensate individuals who are disadvantaged both economically and in terms of expected longevity.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process.

During the preparation of this work the authors used ChatGPT and Refine in order to check grammar, reference consistency, and typewriting. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Auerbach, Alan J., Kerwin Charles, Courtney Coile, William Gale, Dana Goldman, Charles Lucas, Peter Orszag, et al (2017). How the Growing Gap in Life Expectancy May Affect Retirement Benefits and Reforms. *Geneva Papers on Risk and Insurance: Issues and Practice*, 42 (3): 475–499.

Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz (2021). A Comparative Analysis of Gradient Boosting Algorithms. *Artificial Intelligence Review*, 54: 1937–1967.

Bor, Jacob, Gregory H. Cohen, and Sandro Galea (2017). Population Health in an Era of Rising Income Inequality: USA, 1980–2015. *The Lancet*, 389 (10077): 1475–1490.

Boserup, Simon Halphen, Wojciech Kopczuk, and Claus Thustrup Kreiner (2016). The Role of Bequests in Shaping Wealth Inequality: Evidence from Danish Wealth Records. *American Economic Review: Papers and Proceedings*, 106 (5): 656–661.

Boserup, Simon Halphen, Wojciech Kopczuk, and Claus Thustrup Kreiner (2017). Intergenerational Wealth Formation over the Life Cycle: Evidence from Danish Wealth Records 1984-2013. Unpublished working paper.

Brewer, M., B. Etheridge, and C. O'Dea (2017). Why Are Households That Report the Lowest Incomes So Well-Off? *Economic Journal*, 127 (605): F24–F49.

Brown, Denise, Alex D. McMahon, I. David Conway, and Alastair H. Leyland (2019). Measuring Socioeconomic Status and Inequalities. In *Reducing Social Inequalities in Cancer: Evidence and Priorities for Research*, ed. S. Vaccarella et al., IARC Scientific Publications No. 168, Lyon: International Agency for Research on Cancer, ch. 4, pp. 45–56.

Case, Anne, and Angus Deaton (2015). Rising Morbidity and Mortality in Midlife among White Non-Hispanic Americans in the 21st Century. *Proceedings of the National Academy of Sciences*, 112 (49): 15078–15083.

Case, Anne, and Angus Deaton (2020). *Deaths of Despair and the Future of Capitalism.* Princeton, NJ: Princeton University Press.

Case, Anne, and Angus Deaton (2021). Life Expectancy in Adulthood Is Falling for Those without a BA Degree, but as Educational Gaps Have Widened, Racial Gaps Have Narrowed. *Proceedings of the National Academy of Sciences*, 118 (11): e2024777118.

Chetty, Raj, David Grusky, Maximilian Hell, Nathaniel Hendren, Robert Manduca, and Jimmy Narang (2017). The Fading American Dream: Trends in Absolute Income Mobility since 1940. *Science*, 356 (6336): 398–406.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler (2016). The Association between Income and Life Expectancy in the United States, 2001–2014. *JAMA*, 315 (16): 1750–1766.

Christensen, Gunhild T., Drude Molbo, Lars H. Ängquist, Erik L. Mortensen, Kaare Christensen, Thorkild I. A. Sørensen, and Merete Osler (2015). Cohort Profile: The Danish Conscription Database (DCD): A Cohort of 728,160 Men Born from 1939 through 1959. *International Journal of Epidemiology*, 44 (2): 432–440.

Couillard, Benjamin K., Christopher L. Foote, Kavish Gandhi, Ellen Meara, and Jonathan Skinner (2021). Rising Geographic Disparities in US Mortality. *Journal of Economic Perspectives*, 35 (4): 123–146.

Currie, Janet, and Hannes Schwandt (2016). Inequality in Mortality Decreased among the Young while Increasing for Older Adults, 1990–2010. *Science*, 352 (6286): 708–712.

Dahl, Gordon B., Claus Thustrup Kreiner, Torben Heien Nielsen, and Benjamin Ly Serena (2024). Understanding the Rise in Life Expectancy Inequality. *Review of Economics and Statistics*, 106 (2): 566–575.

Danesh, Kaveh, Jonathan T. Kolstad, Johannes Spinnewijn, and William D. Parker (2024). The Chronic Disease Index: Analyzing Health Inequalities over the Lifecycle. *National Bureau of Economic Research Working Paper* No. 32577.

Deaton, Angus (2013). *The Great Escape: Health, Wealth, and the Origins of Inequality.* Princeton, NJ: Princeton University Press.

Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer (2018). Predictive Modeling of U.S. Health Care Spending in Late Life. *Science*, 360 (6396): 1462–1465.

Frazier, Peter I. (2018). A Tutorial on Bayesian Optimization. arXiv preprint arXiv:1807.02811.

Ganna, Andrea, and Erik Ingelsson (2015). 5-Year Mortality Predictors in 498,103 UK Biobank Participants: A Prospective Population-Based Study. *The Lancet*, 386 (9993): 533–540.

Gavrilov, Leonid, and Natalia Gavrilova (1991). *The Biology of Life Span: A Quantitative Approach.* New York: Harwood Academic Publishers.

Gavrilov, Leonid, and Natalia Gavrilova (2019). New Trend in Old-Age Mortality: Gompertzialization of Mortality Trajectory. *Gerontology*, 65 (5): 451–457.

Glei, Dana A., Chioun Lee, and Maxine Weinstein (2022). Assessment of Mortality Disparities by Wealth Relative to Other Measures of Socioeconomic Status Among US Adults. JAMA Network Open, 5 (4): e226547.

Gompertz, Benjamin (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London*, 115: 513–583.

Green, Anders (1996). The Danish Conscription Registry: A Resource for Epidemiological Research. *Danish Medical Bulletin*, 43 (5): 464–467.

Hagen, Johannes, Lisa Laun, Charlotte Lucke, and Mårten Palme (2025). The rising income gradient in life expectancy in Sweden over six decades. *Proceedings of the National Academy of Sciences of the United States of America*, 122 (14): e2418145122.

Hederos, K., M. Jäntti, L. Lindahl, and J. Torssander (2018). Trends in Life Expectancy by Income and the Role of Specific Causes of Death. *Economica*, 85: 606–625.

Hendi, Arun S. (2017). Trends in Education-Specific Life Expectancy, Data Quality, and Shifting Education Distributions: A Note on Recent Research. Demography, 54 (3): 1203–1213.

Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger (2018). DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. *BMC Medical Research Methodology*, 18: 24.

Kinge, Jonas Minet, Jørgen Heibø Modalsli, Simon Øverland, Håkon Kristian Gjessing, Mette Christophersen Tollånes, Ann Kristin Knudsen, Vegard Skirbekk, Bjørn Heine Strand, Siri Eldevik Håberg, and Stein Emil Vollset (2019). Association of Household Income with Life Expectancy and Cause-Specific Mortality in Norway, 2005–2015. *JAMA*, 321 (19): 1916–1925.

Kreiner, Claus Thustrup, Torben Heien Nielsen, and Benjamin Ly Serena (2018). Role of Income Mobility for the Measurement of Inequality in Life Expectancy. *Proceedings of the National Academy of Sciences*, 115 (46): 11754–11759.

Leth-Petersen, Søren (2010). Intertemporal Consumption and Credit Constraints: Does Total Expenditure Respond to an Exogenous Shock to Credit? *American Economic Review*, 100 (3): 1080–1103.

Mackenbach, Johan P., Irina Stirbu, Albert-Jan R. Roskam, Maartje M. Schaap, Gwenn Menvielle, Mall Leinsalu, and Anton E. Kunst (2008). Socioeconomic Inequalities in Health in 22 European Countries. *New England Journal of Medicine*, 358 (23): 2468–2481.

Mackenbach, Johan P., José Rubio Valverde, Barbara Artnik, Matthias Bopp, Henrik Brønnum-Hansen, Patrick Deboosere, Ramune Kalediene, et al (2018). Trends in Health Inequalities in 27 European Countries. *Proceedings of the National Academy of Sciences*, 115 (25): 6440–6445.

Meara, E. R., S. Richards, and D. M. Cutler (2008). The Gap Gets Bigger: Changes in Mortality and Life Expectancy, by Education, 1981–2000. *Health Affairs*, 27 (2): 350–360.

Milligan, Kevin (2024). The Time of Your Life: The Mortality and Longevity of Canadians. *National Bureau of Economic Research Working Paper* No. 33066, October.

Milligan, Kevin, and Tammy Schirle (2021). The Evolution of Longevity: Evidence from Canada. *Canadian Journal of Economics/Revue canadienne d'économique*, 54: 164–192.

Mortensen, Erik L., June M. Reinisch, and Thomas W. Teasdale (1989). Intelligence as Measured by the WAIS and a Military Draft Board Group Test. *Scandinavian Journal of Psychology*, 30 (4): 315–318.

National Academies of Sciences, Engineering, and Medicine (2015). *The Growing Gap in Life Expectancy by Income: Implications for Federal Programs and Policy Responses.* Washington, DC: The National Academies Press.

OECD (2022). Educational Inequalities in Longevity among OECD Countries around 2016. In *OECD Papers on Well-Being and Inequalities*, No. 8. Paris: OECD Publishing.

OECD/European Union (2018). Inequalities in Life Expectancy. In *Health at a Glance: Europe 2018: State of Health in the EU Cycle.* Paris/Brussels: OECD Publishing/European Union.

Piketty, Thomas, and Emmanuel Saez (2014). Inequality in the Long Run. *Science*, 344 (6186): 838–843.

Puterman, Eli, Jordan Weiss, Benjamin A. Hives, Alison Gemmill, Deborah Karasek, Wendy Berry Mendes, and David H. Rehkopf (2020). Predicting Mortality from 57 Economic, Behavioral, Social, and Psychological Factors. *Proceedings of the National Academy of Sciences*, 117 (30): 16273–16282.

United Nations (2022). On the Importance of Monitoring Inequality in Life Expectancy. *UN DESA Policy Brief* No. 145, 16 December 2022.

United Nations (2023). *World Social Report 2023: Leaving No One Behind in an Ageing World.* New York: United Nations Department of Economic and Social Affairs.

van Raalte, Alyson A., Anton E. Kunst, Olle Lundberg, Mall Leinsalu, Pekka Martikainen, Barbara Artnik, Patrick Deboosere, Irina Stirbu, Bogdan Wojtyniak, and Johan P. Mackenbach (2012). The Contribution of Educational Inequalities to Lifespan Variation. *Population Health Metrics*, 10: 3 (1–10).

Online Appendix
Socioeconomic Inequality in Longevity: A Multidimensional Approach

Paul Bingley, Claus Thustrup Kreiner, and Benjamin Ly Serena

Additional details on socioeconomic (SES) indicators

Below we describe the specific construction of each SES indicator. Variables in ALL CAPS signify that they refer to variable names in the raw data as used by Statistics Denmark.

Income

We measure income net of universal transfers in line with previous work (Kreiner et al., 2018; Dahl et al., 2024):

$$\text{Income} = \text{PERINDKIALT} - \text{QPENSNY} - \text{KONTHJ} - \text{ANDOVERFORSEL} \tag{5}$$

PERINDKIALT is total income, QPENSNY is public pensions (disability insurance benefits, public retirement pension), KONTHJ is cash assistance, and ANDOVERFORSEL is other universal public transfers.

Education

We measure education as the length of education in years and dummies for individual educational qualifications (e.g., engineer, economist, etc.) based on the variable HFAUDD.

Wealth

We measure wealth as the difference between assets (QAKTIVF) and liabilities (QPASSIV). Assets include third-party reported housing assets, bank account balances, stocks and bonds, equity in private companies, and self-reported values of cars, boats, and campers. The self-reported components of recorded assets are negligible (see e.g., Appendix Figure A1 in Boserup et al. (2017)).

Liabilities include bank loans, mortgages, student debt, and debt to the tax authorities. Housing assets are known to be undervalued by tax authorities, while housing liabilities (mortgages, etc.) are based on the actual price of buying the house. Therefore, we follow previous research (Boserup et al., 2016; Leth-Petersen, 2010) and scale up the tax-assessed value of housing assets by 20% to match average actual house prices.

During our analysis period, all wealth is recorded in the husband's name for married couples. Therefore, we compute the total wealth of married couples and divide it equally between the two partners.

IQ

At age 18, all Danish men are requested to attend a regional conscription board to assess their fitness for military service. The Danish Conscription Database contains information on these assessments (Christensen et al., 2015). Our IQ measure is the Danish Armed Forces Qualification Test, called the Børg Prien Prøve, which was developed for military recruitment to measure cognitive ability. The test contains 78 items: 19 letter matrices, 24 verbal analogies, 17 number series, and 18 geometric figures, to be completed in 45 minutes. The score is the total number of correct responses. The correlation coefficient between the Børg Prien Prøve scores and the Wechsler Adult Intelligence Scale is 0.82 (Mortensen et al., 1989). While assessments can be postponed to age 26, 77% of men were tested at ages 18-20.

Between five and ten percent of men were deemed unfit for service and exempt from assessment because of medical conditions, e.g., intellectual disability or epilepsy (Green, 1996). Predating the civil registration system, records were identified by names and dates of birth written on cards. Two percent of records could not be matched to administrative registers (Christensen et al., 2015).

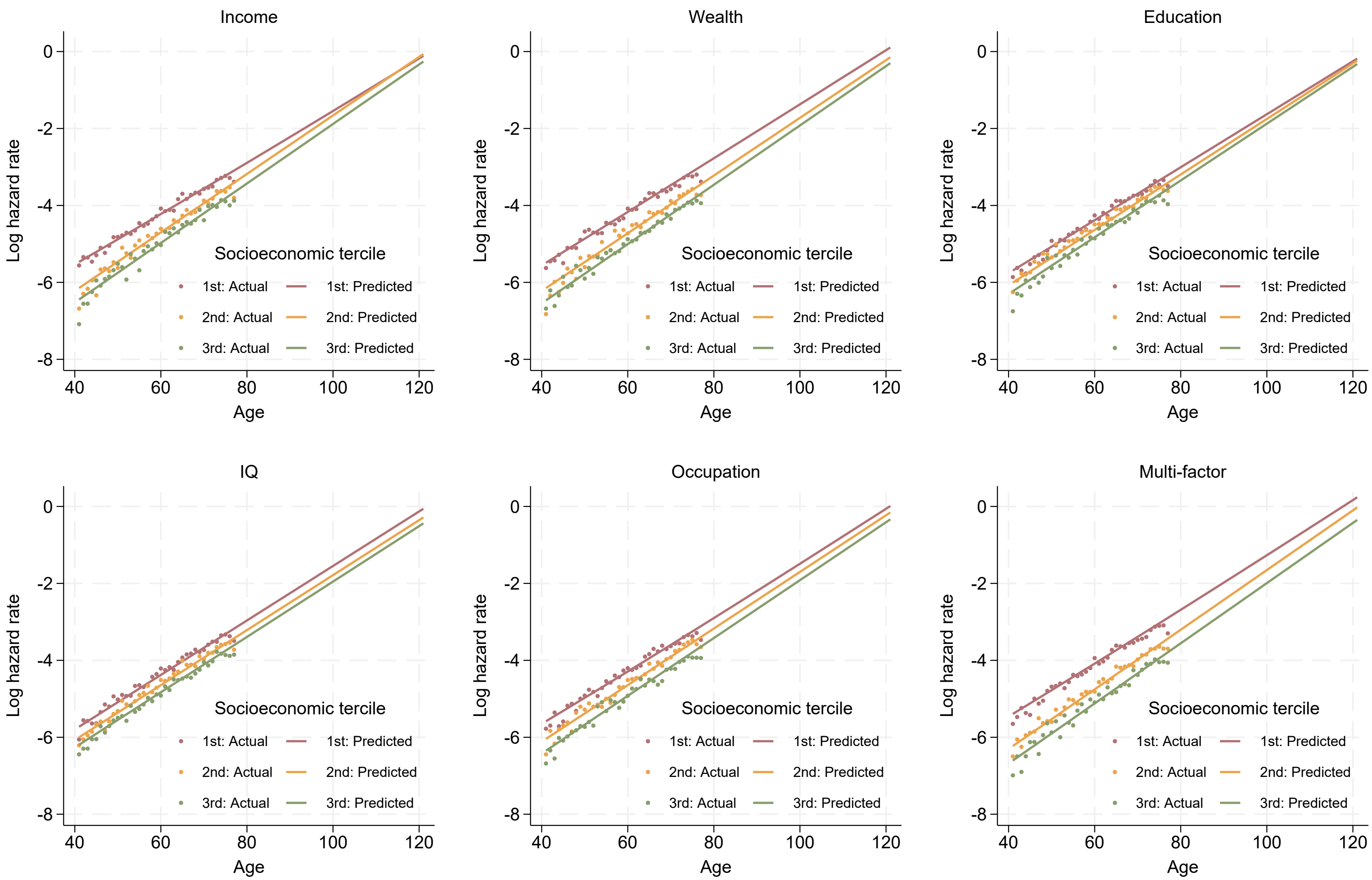
Occupation

We measure occupation by including dummy variables for occupation codes, using the variable NYSTGR based on a classification similar to the Disco88 coding system. Additionally, we include dummies for broader occupational categories derived from the first digit of this variable. For instance, these categories include the self-employed with fewer than 20 employees.

In Figure A2 below, we plot predicted life expectancy against continuous measures of SES. To convert occupational categories into a continuous measure suitable for plotting against life expectancy, we employ the commonly used International Socio-Economic Index of Occupational Status (Ganzeboom et al., 1992), which ranks occupations according to social prestige.

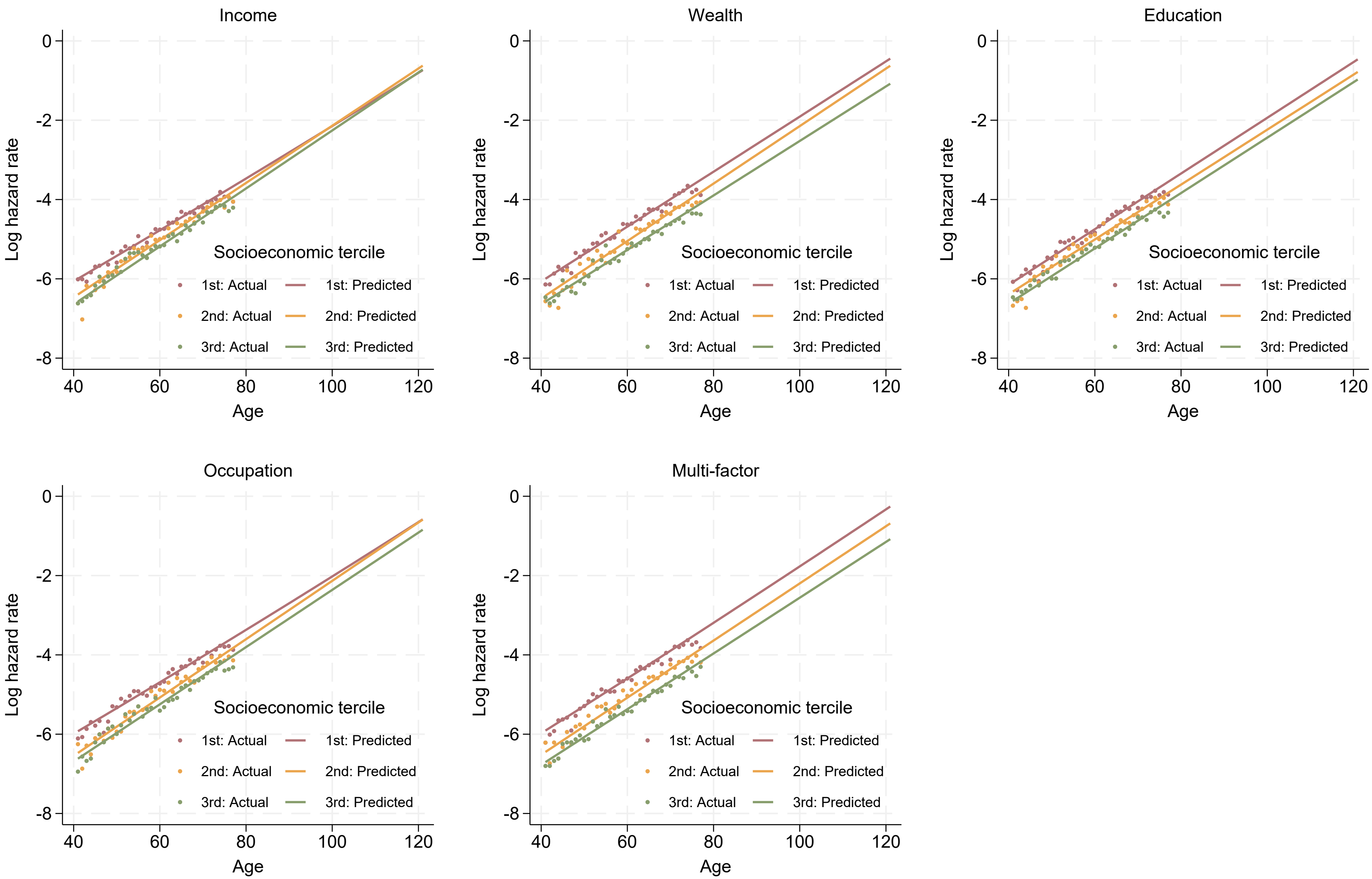
Appendix Figures and Tables

Figure A1: Gompertz approximations by age and SES for men



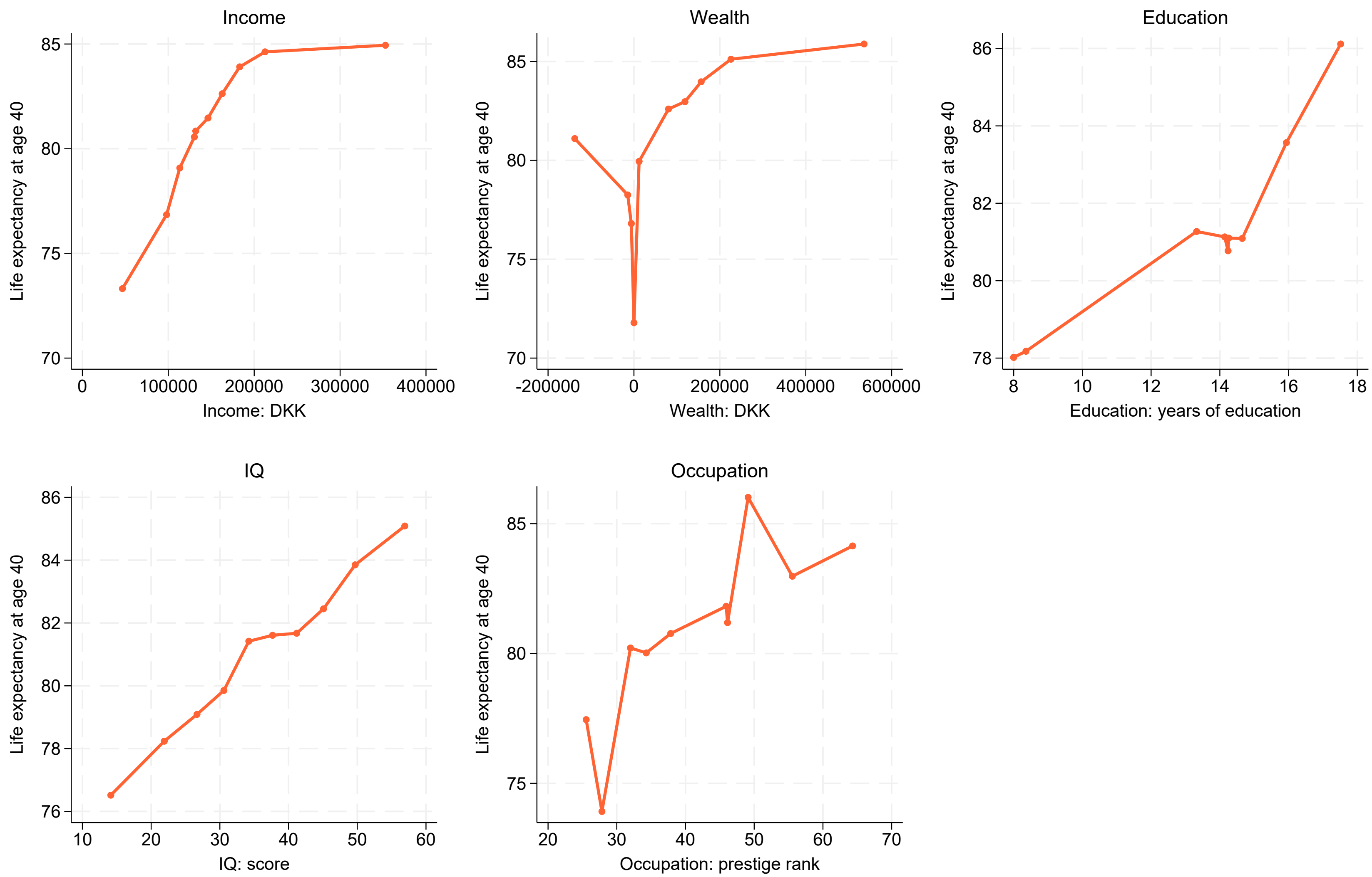
Note: The figure plots predicted and actual log mortality hazard rates by age and SES tercile. Terciles are constructed from predicted life expectancy based on the training data and the Gompertz model. The lines represent predicted hazard rates based on the training data. The dots represent Kaplan-Meier estimates of actual log hazard rates in the test sample.

Figure A1: (Continued). Gompertz approximations by age and SES for women



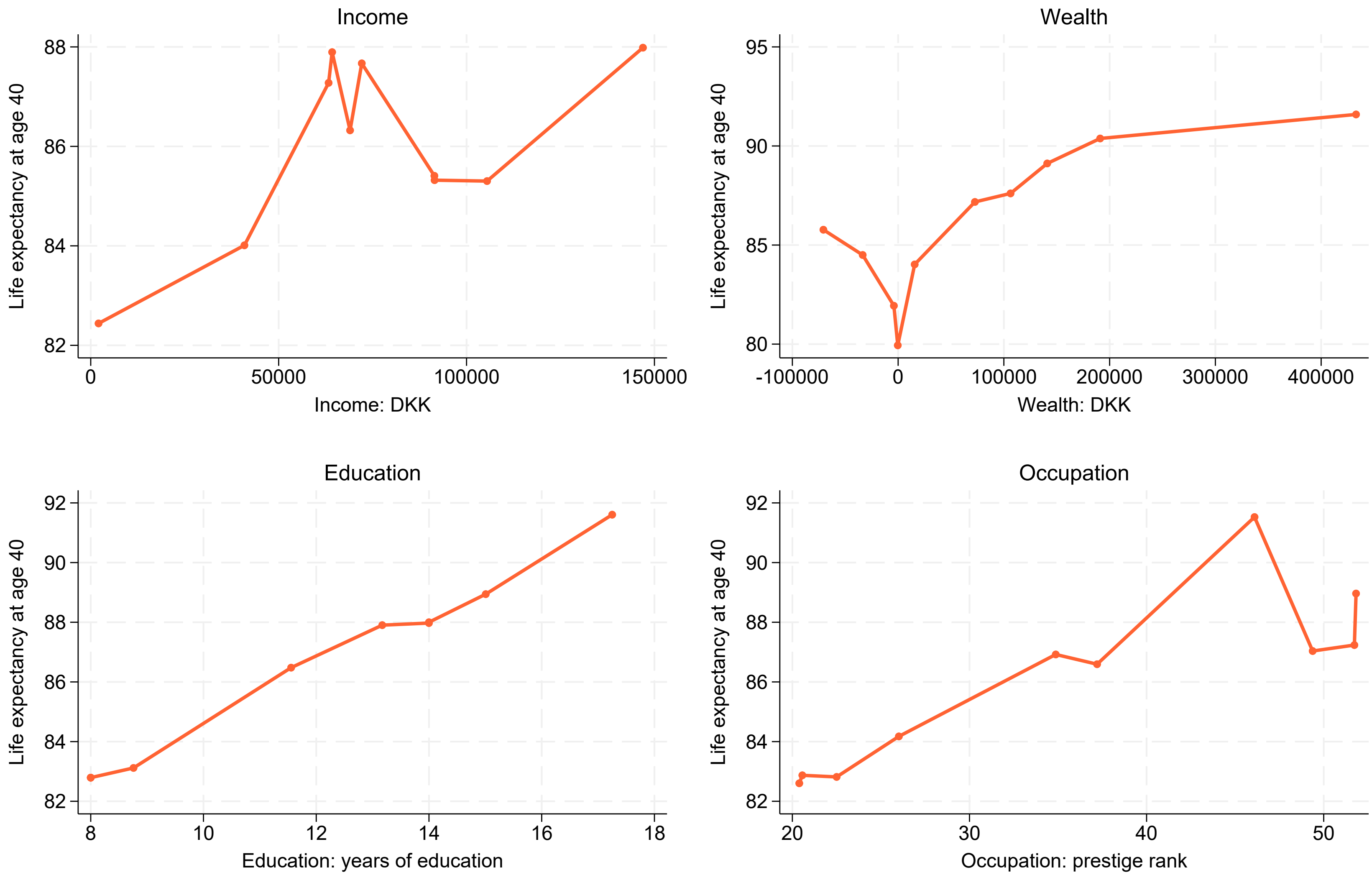
Note: The figure plots predicted and actual log mortality hazard rates by age and SES tercile. Terciles are constructed from predicted life expectancy based on the training data and the Gompertz model. The lines represent predicted hazard rates based on the training data. The dots represent Kaplan-Meier estimates of actual log hazard rates in the test sample.

Figure A2: Association between life expectancy and absolute levels of socioeconomic indicators for men



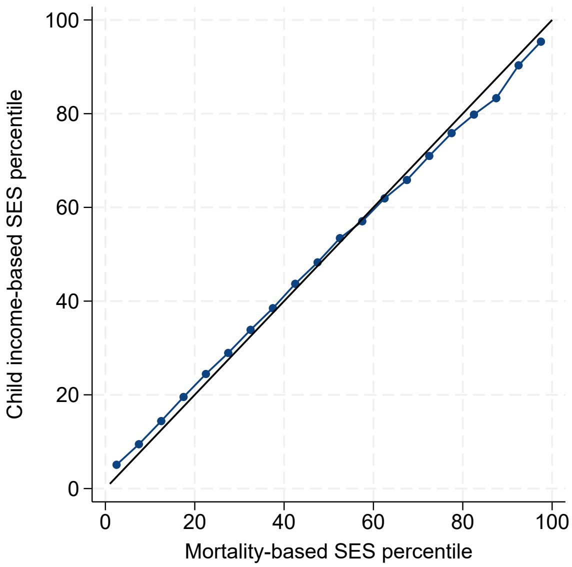
Note: The figure plots deciles of predicted life expectancy against the within-decile mean of the SES indicators used to make the prediction. The SES deciles are constructed from predicted life expectancy based on the training data and the Gompertz model.

Figure A2: (Continued). Association between life expectancy and absolute levels of socioeconomic indicators for women

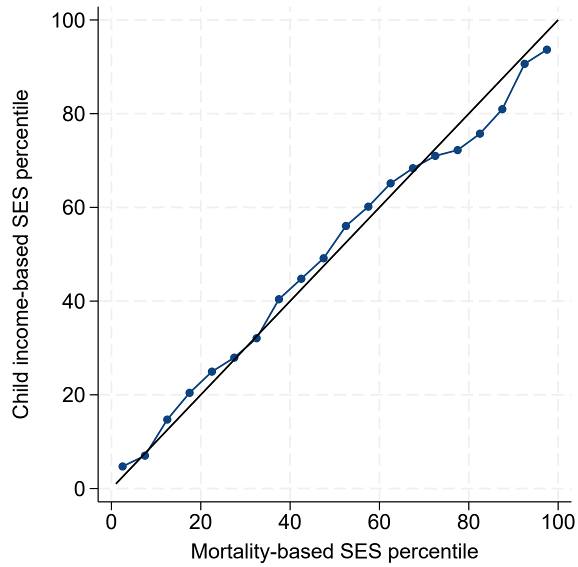


Note: The figure plots deciles of predicted life expectancy against the within-decile mean of the SES indicators used to make the prediction. The SES deciles are constructed from predicted life expectancy based on the training data and the Gompertz model.

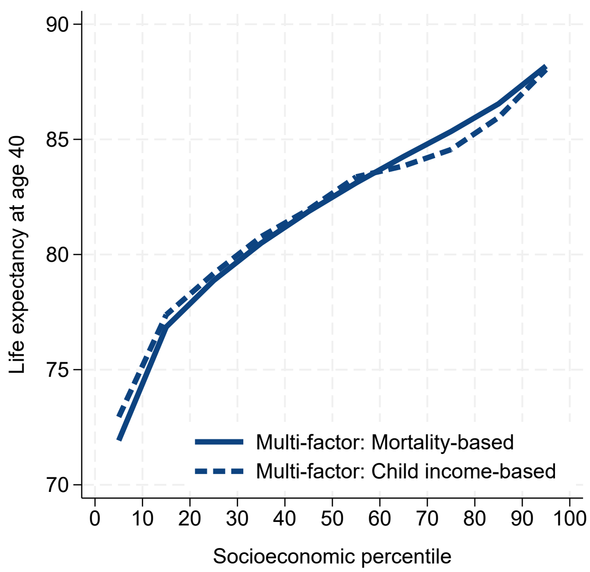
Figure A3: Results using predictions of children's adult income rank to form SES ranks



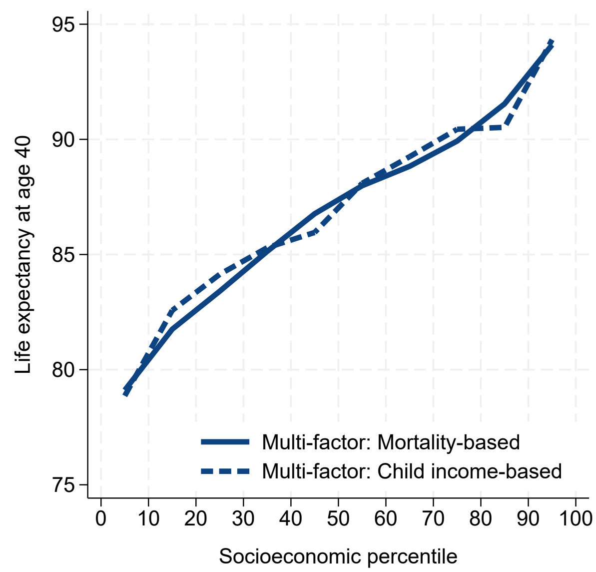
(a) Ranking comparison, Men



(b) Ranking comparison, Women



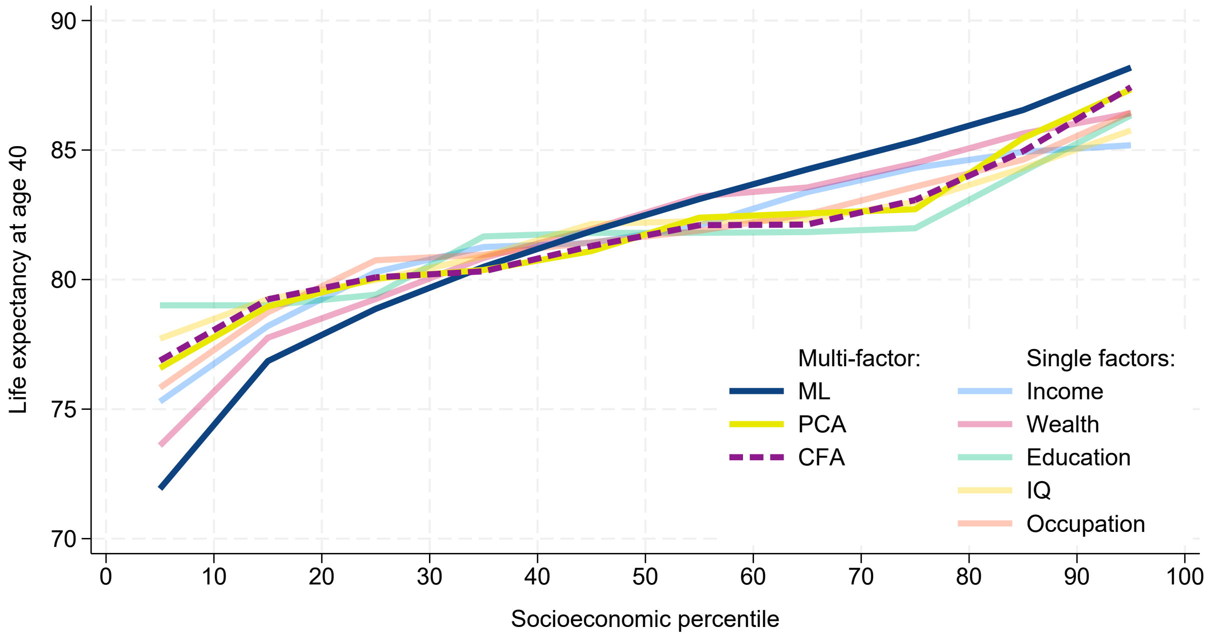
(c) Longevity gradient comparison, Men



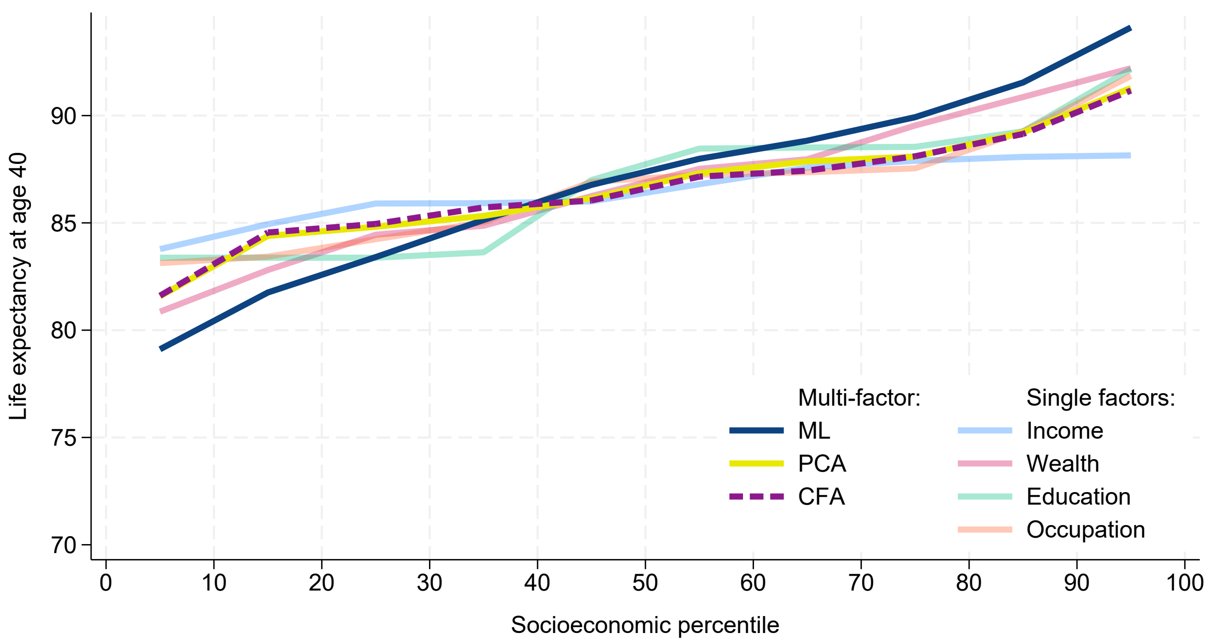
(d) Longevity gradient comparison, Women

Note: The figure shows results using a multi-factor SES rank based on machine-learning predictions of the adult income ranks of children of individuals in our main sample. We compare these ranks with those from our standard approach, which predicts the future mortality of the individuals themselves. Children's income rank is measured as their within-year income rank at age 38—the same age at which parental SES is measured—and averaged across children.

Figure A4: Results using Common Factor Analysis and Principal Component Analysis to form SES ranks



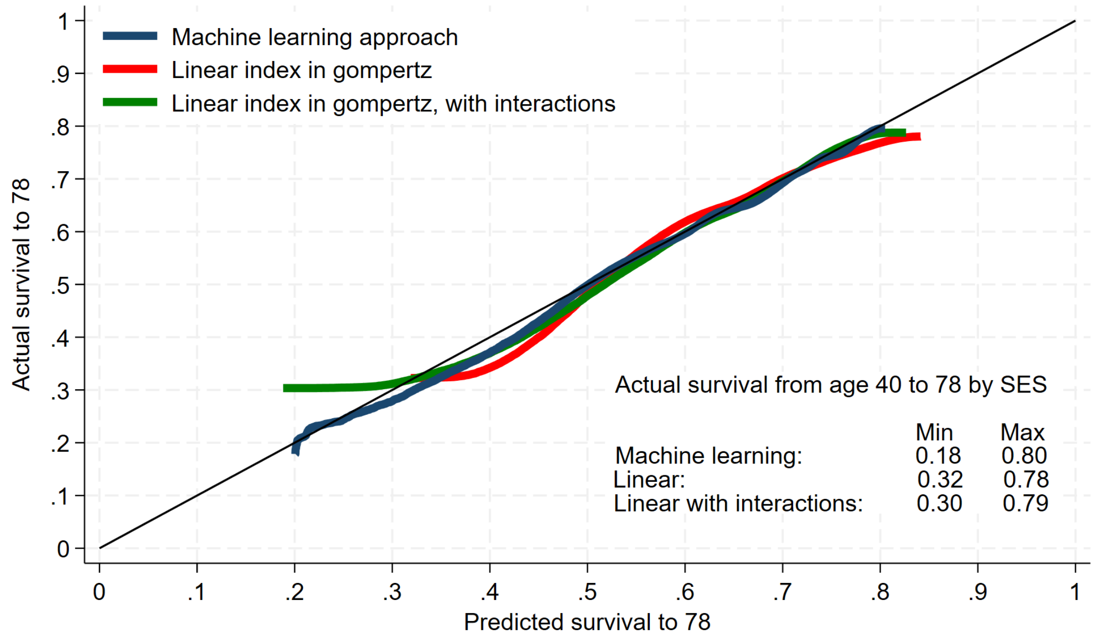
(a) Men



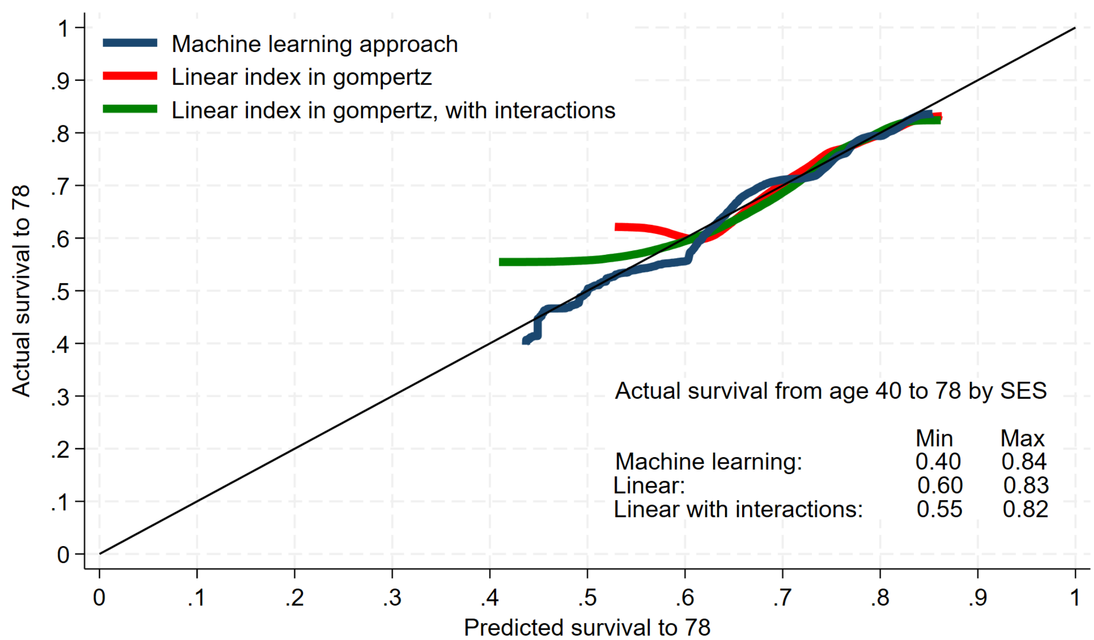
(b) Women

Note: The figure shows longevity gradients in SES using the first components of Common Factor Analysis (CFA) and Principal Component Analysis (PCA) to form multi-factor SES ranks. For each method, we include the five SES indicators as continuous variables (length of education for education, occupational prestige for occupation), ranked within cohort and standardized. We compare the CFA and PCA results with our baseline multi-factor and single-factor gradients.

Figure A5: Results using simple parametric alternatives to the machine-learning approach



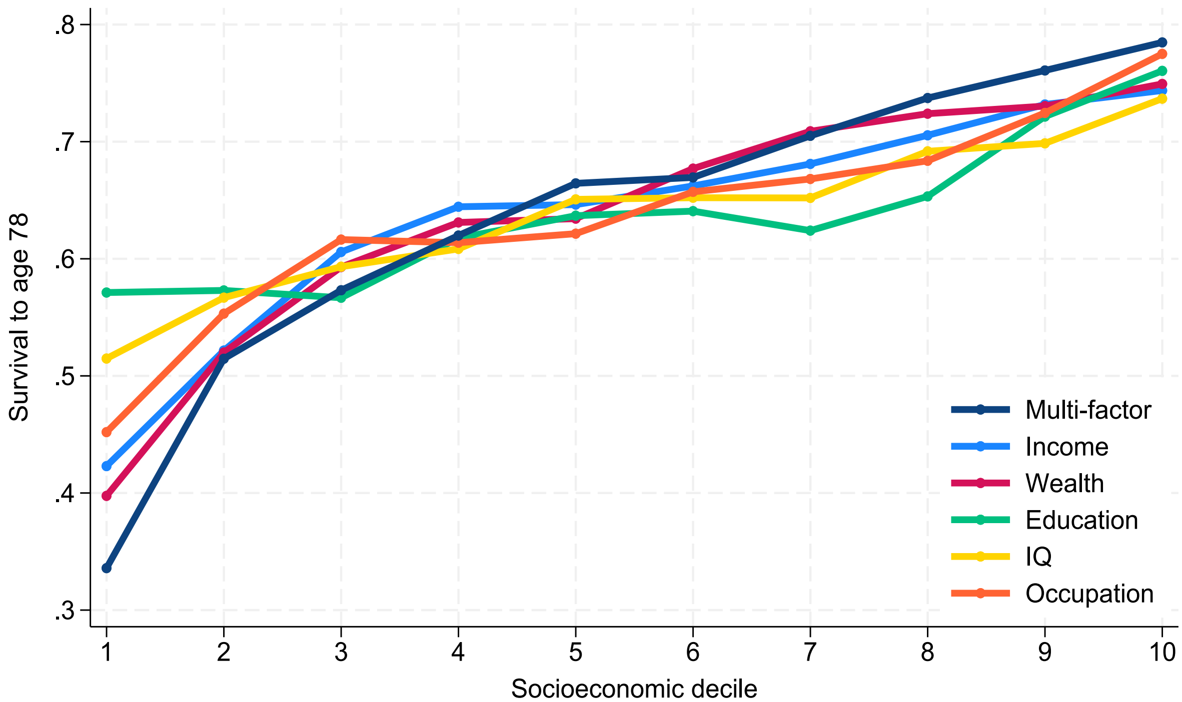
(a) Men



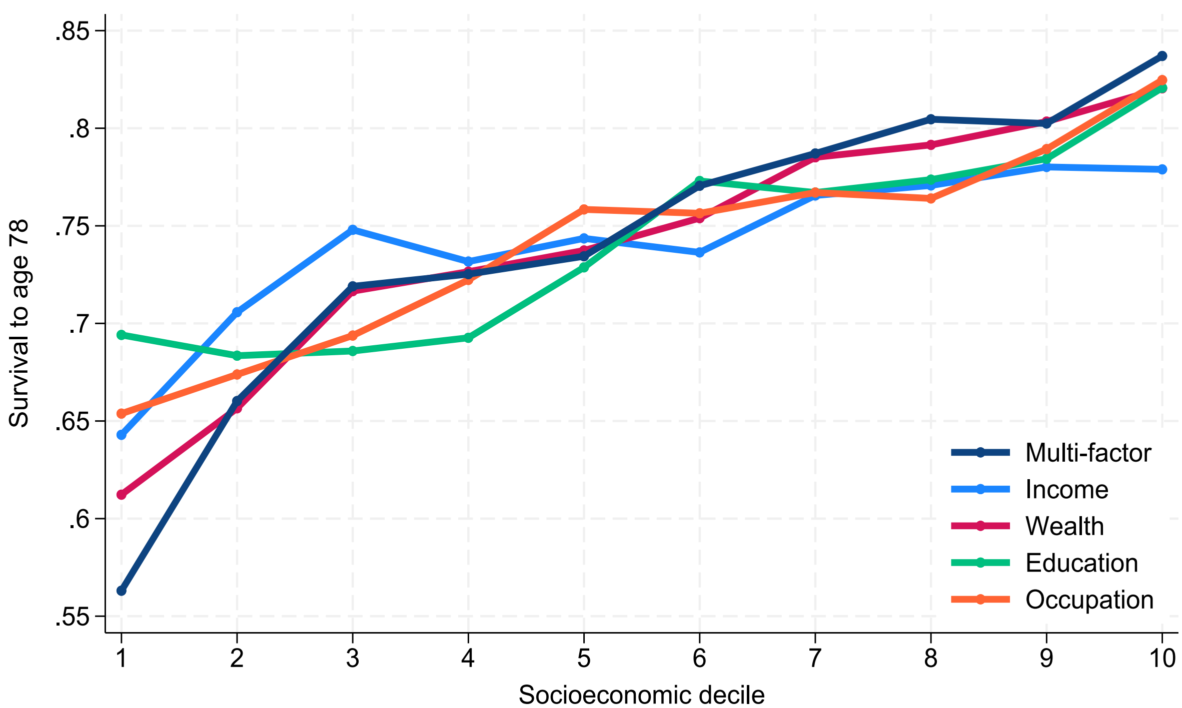
(b) Women

Note: The figure shows actual versus predicted survival from age 40 to 78, based on different ways of combining SES indicators. The horizontal axis shows the range of predicted survival probabilities generated by each method, while the vertical axis shows actual survival among individuals with the corresponding predicted survival probability, estimated nonparametrically using local linear regression. The blue line shows the baseline machine-learning approach. The red and green lines show two parametric alternatives based on a Gompertz survival model, $\log(\lambda) = \alpha x + \beta(x) \cdot age$, where x includes continuous percentile ranks of the five SES indicators: income, wealth, education length in years, IQ score (for men), and occupational prestige associated with a given occupation code. The green specification additionally includes full interactions between the SES indicators. The reported minimum and maximum values summarize the range of predicted survival probabilities generated by each approach. Bandwidths for the local linear regressions are selected using a rule-of-thumb procedure.

Figure A6: Actual survival from age 40 to 78



(a) Men



(b) Women

Note: The figure plots the actual share of individuals surviving in the test sample from age 40 to 78 by SES deciles. Deciles are constructed from predicted life expectancy based on the training data and the Gompertz model.

Table A1: Overlap between predicted multi-factor life-expectancy groups and SES rankings

***A** Top 10%: Share in top 10% of predicted life expectancy also in top 10% of SES rankings*

	Income (1)	Wealth (2)	IQ (3)	Education (4)	Occupation (5)	Average (1)–(5)	Ranking based on average position across indicators
Men	38	36	41	49	41	**41**	**79**
Women	25	37	–	62	47	**42**	**72**

***B** Top 1%: Share in top 1% of predicted life expectancy also in top 1% of SES rankings*

	Income (1)	Wealth (2)	IQ (3)	Education (4)	Occupation (5)	Average (1)–(5)	Ranking based on average position across indicators
Men	4	6	17	8	13	**9**	**62**
Women	0	6	–	7	3	**4**	**15**

***C** Bottom 10%: Share in bottom 10% of predicted life expectancy also in bottom 10% of SES rankings*

	Income (1)	Wealth (2)	IQ (3)	Education (4)	Occupation (5)	Average (1)–(5)	Ranking based on average position across indicators
Men	53	62	30	22	44	**42**	**73**
Women	37	47	–	25	19	**32**	**81**

Note: The table reports the percentage of individuals in the top 10%, top 1%, or bottom 10% of predicted life expectancy based on the multi-factor model who are also in the corresponding group of each SES ranking. "Average (1)–(5)" reports the average overlap across the individual SES indicators. The final column ranks individuals by their average position across the SES indicators and reports the resulting overlap.

Table A2: Life expectancy inequality without sample restrictions

	Multi-factor	Education	Income	Wealth	IQ	Occupation
Men						
Min-Max	65 - 91	78 - 86	69 - 86	71 - 86	76 - 86	69 - 89
P10-P90	75 - 87	78 - 85	77 - 86	77 - 86	78 - 84	77 - 85
Linear gradient	0.173	0.086	0.118	0.135	0.078	0.117
SES share (%) of var.	11.6	2.9	5.9	7.5	2.3	6.1
Women						
Min-Max	75 - 98	84 - 93	81 - 89	81 - 92		83 - 93
P10-P90	81 - 92	84 - 91	84 - 89	81 - 92		84 - 91
Linear gradient	0.150	0.099	0.062	0.114		0.090
SES share (%) of var.	7.7	3.7	1.4	4.3		3.0

Note: The table reports four different statistics for the distribution of predicted life expectancy at age 40 in the test sample without imposing sample restrictions. The results are based on all individuals in the cohorts 1942-1944 who are alive at age 40. The first row reports the minimum and maximum; the second row reports the 10^{th} and 90^{th} percentiles; the third row reports the slope of a linear regression line across percentiles of the SES measure; the fourth row reports the percent of the variation in life expectancy accounted for by SES. The column headers show which single SES indicator the prediction uses. Multi-factor approach refers to the combination of all indicators.

Table A3: Life expectancy inequality for men without IQ restriction

	Multi-factor	Education	Income	Wealth	Occupation
Min-Max	66 - 89	77 - 87	67 - 85	69 - 86	67 - 86
P10-P90	73 - 86	77 - 85	75 - 84	76 - 85	76 - 86
Linear gradient	0.176	0.088	0.121	0.139	0.121
SES share (%) of var.	12.3	3.2	6.6	8.1	6.7

Note: The table reports four different statistics for the distribution of predicted life expectancy at age 40 in the test sample without imposing non-missing IQ information for men. The first row reports the minimum and maximum; the second row reports the 10^{th} and 90^{th} percentiles; the third row reports the slope of a linear regression line across percentiles of the SES measure; the fourth row reports the percent of the variation in life expectancy accounted for by SES. The column headers show which single SES indicator the prediction uses. Multi-factor approach refers to the combination of all indicators.

Table A4: Life expectancy inequality, using household averages of socioeconomic indicators

	Multi-factor	Education	Income	Wealth	Occupation
Men					
Min-Max	65 - 89	76 - 87	67 - 84	69 - 85	68 - 89
P10-P90	73 - 86	76 - 83	79 - 84	76 - 85	76 - 85
Linear gradient	0.172	0.095	0.091	0.138	0.122
SES share (%) of var.	11.8	3.6	4.4	8.0	6.6
Women					
Min-Max	73 - 97	78 - 94	75 - 90	80 - 92	75 - 94
P10-P90	80 - 92	83 - 90	83 - 90	81 - 91	82 - 90
Linear gradient	0.168	0.123	0.083	0.125	0.116
SES share (%) of var.	9.4	5.1	3.2	5.0	5.0

Note: The table reports four different statistics for the distribution of predicted life expectancy at age 40 in the test sample. The first row reports the minimum and maximum; the second row reports the 10^{th} and 90^{th} percentiles; the third row reports the slope of a linear regression line across percentiles of the SES measure; the fourth row reports the percent of the variation in life expectancy accounted for by SES. The column headers show which single SES indicator the prediction uses. Multi-factor approach refers to the use of all indicators in the prediction. Continuous indicators are measured as the average within the household. For categorical variables, we include dummies for each category for both the main individual and the spouse. For unmarried individuals, we code the spouse-related dummy variables as missing values.

Table A5: Life expectancy inequality, excluding recipients of disability insurance benefits at age 38

	Multi-factor	Education	Income	Wealth	IQ	Occupation
Men						
Min-Max	65 - 88	78 - 86	72 - 85	70 - 86	76 - 86	69 - 87
P10-P90	74 - 87	78 - 85	76 - 85	76 - 86	78 - 84	77 - 85
Linear gradient	0.158	0.072	0.102	0.132	0.080	0.102
SES share (%) of var.	10.2	2.3	4.3	7.3	2.5	4.4
Women						
Min-Max	75 - 96	83 - 92	85 - 89	81 - 93		82 - 94
P10-P90	81 - 92	84 - 92	85 - 88	81 - 90		83 - 91
Linear gradient	0.143	0.097	0.041	0.122		0.092
SES share (%) of var.	7.2	3.6	0.6	5.2		3.2

Note: The table reports four different statistics for the distribution of predicted life expectancy at age 40 in the test sample. The first row reports the minimum and maximum; the second row reports the 10^{th} and 90^{th} percentiles; the third row reports the slope of a linear regression line across percentiles of the SES measure; the fourth row reports the percent of the variation in life expectancy accounted for by SES. The column headers show which single SES indicator the prediction uses. Multi-factor approach refers to the use of all indicators in the prediction.

Table A6: Life expectancy inequality, using Kannisto model for extrapolations

	Multi-factor	Education	Income	Wealth	IQ	Occupation
Men						
Min-Max	63 - 88	78 - 87	67 - 85	68 - 85	76 - 85	67 - 87
P10-P90	74 - 86	78 - 86	75 - 85	76 - 84	77 - 84	77 - 85
Linear gradient	0.161	0.086	0.122	0.128	0.085	0.105
SES share (%) of var.	10.6	3.2	6.3	7.4	2.7	5.3
Women						
Min-Max	75 - 97	83 - 92	81 - 88	80 - 90		83 - 93
P10-P90	80 - 93	83 - 91	83 - 88	80 - 90		83 - 90
Linear gradient	0.167	0.103	0.053	0.122		0.085
SES share (%) of var.	9.2	3.8	1.1	5.0		2.7

Note: The table reports four different statistics for the distribution of predicted life expectancy at age 40 in the test sample. The first row reports the minimum and maximum; the second row reports the 10^{th} and 90^{th} percentiles; the third row reports the slope of a linear regression line across percentiles of the SES measure; the fourth row reports the percent of the variation in life expectancy accounted for by SES. The column headers show which single SES indicator the prediction uses. Multi-factor approach refers to the use of all indicators in the prediction. The Kannisto model differs from the Gompertz model by imposing a plateau in mortality rates at old ages (Gavrilov and Gavrilova, 2019). This difference is apparent from how they model the relationship between the hazard rate and age:
Gompertz: $\lambda(age) = \alpha x e^{\beta x \cdot age}$. Kannisto: $\lambda(age) = \frac{\alpha x e^{\beta x \cdot age}}{1+\alpha x e^{\beta x \cdot age}}$.

References

Boserup, S. H., Kopczuk, W., and Kreiner, C. T. (2016). The role of bequests in shaping wealth inequality: Evidence from Danish wealth records. *American Economic Review, Papers and Proceedings*, 106(5):656–661.

Boserup, S. H., Kopczuk, W., and Kreiner, C. T. (2017). Intergenerational wealth formation over the life cycle: Evidence from Danish wealth records 1984–2013. Unpublished working paper.

Christensen, G. T., Molbo, D., Ängquist, L. H., Mortensen, E. L., Christensen, K., Sørensen, T. I. A., and Osler, M. (2015). Cohort profile: The Danish Conscription Database (DCD): A cohort of 728,160 men born from 1939 through 1959. *International Journal of Epidemiology*, 44(2):432–440.

Dahl, G. B., Kreiner, C. T., Nielsen, T. H., and Serena, B. L. (2024). Understanding the rise in life expectancy inequality. *Review of Economics and Statistics*, 106(2):566–575.

Ganzeboom, H. B., De Graaf, P. M., and Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 21(1):1–56.

Gavrilov, L. R. and Gavrilova, N. (2019). New trend in old-age mortality: Gompertzialization of mortality trajectory. *Gerontology*, 65(5):451–457.

Green, A. (1996). The Danish Conscription Registry: a resource for epidemiological research. *Danish medical bulletin*, 43(5):464—467.

Kreiner, C. T., Nielsen, T. H., and Serena, B. L. (2018). Role of income mobility for the measurement of inequality in life expectancy. *Proceedings of the National Academy of Sciences of the United States of America*, 115(46):11754–11759.

Leth-Petersen, S. (2010). Intertemporal consumption and credit constraints: Does total expenditure respond to an exogenous shock to credit? *American Economic Review*, 100(3):1080–1103.

Mortensen, E. L., Reinisch, J. M., and Teasdale, T. W. (1989). Intelligence as measured by the WAIS and a military draft board group test. *Scandinavian Journal of Psychology*, 30(4):315–318.