CENTER FOR
ECONOMIC
BEHAVIOR &
INEQUALITY

CEBI WORKING PAPER SERIES

Working Paper 14/25

SOCIOECONOMIC INEQUALITY IN LONGEVITY: A MULTIDIMENSIONAL APPROACH

Paul Bingley

Claus Thustrup Kreiner

Benjamin Ly Serena

ISSN 2596-447X

CEBI
Department of Economics
University of Copenhagen
www.cebi.ku.dk

Socioeconomic Inequality in Longevity: A Multidimensional Approach*

Paul Bingley[a]

Claus Thustrup Kreiner[b]

Benjamin Ly Serena[c]

November 2025

Abstract

Socioeconomic inequality in longevity is typically measured using a single socioeconomic indicator such as education or income. We combine multiple indicators—education, income, occupation, wealth, and IQ scores—and apply machine learning to measure inequality in longevity. Using Danish population-wide data spanning 40 years, we track mortality for the 1942–44 birth cohorts from age 40 onwards to estimate life expectancy by socioeconomic status. Individuals at the top of the socioeconomic distribution live nearly 25 years longer than those at the bottom. The socioeconomic gradient in life expectancy becomes 50–150% steeper when using multiple indicators.

Keywords: Life Expectancy, Inequality, Machine Learning

JEL Classification: I14

---

* We are grateful to James Banks, Thomas Høgholm Jørgensen, and participants at the Nordic Public Policy Symposium 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, the gradient between life expectancy and latent SES may be underestimated when SES is proxied by only one indicator. If all SES indicators capture the same variation between latent SES and life expectancy, combining them will not affect the estimated gradient. In the opposite extreme case, with uncorrelated SES indicators, the combined gradient would equal the sum of the single-factor gradients. Where real-world data fall between these two extremes remains unknown.

In this paper, we combine multiple SES indicators using machine learning to construct a comprehensive measure of the SES gradient in life expectancy.

The advantage of using machine learning is its ability to capture interaction effects and flexibly combine multiple SES indicators in a data-driven way while avoiding overfitting. In other words, it allows us to exploit rich, multi-dimensional SES data without the model mistaking spurious correlations in the estimation sample for meaningful predictors of life expectancy.

---

[2] Deaton 2013; NAS 2015; Case & Deaton 2015, 2020; Currie & Schwandt 2016; Chetty et al. 2016; Auerbach et al. 2017; Bor, Cohen & Galea 2017; Kreiner, Nielsen & Serena 2018; Mackenbach et al. 2018; Couillard et al. 2021; Dahl et al. 2024; Danesh et al. 2024; Hagen et al. 2025.

[3] Meara, Richards & Cutler 2008; Mackenbach et al. 2008, 2018; Chetty et al. 2016; Kreiner, Nielsen & Serena 2018; OECD 2018; Hederos et al. 2018; Kinge et al. 2019; Case & Deaton 2021; Milligan & Schirle 2021; Dahl et al. 2024; Milligan 2024.

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. Beyond this stark contrast, the gradient across the distribution implies that moving up ten percentiles is associated with an average increase in life expectancy of 1.6-1.7 years. Finally, we analyze how much of the variance in life expectancy can be explained by SES. We find that SES accounts for 12% of the variance for men and 8% for women.

These findings contribute to the literature (op.cit.) measuring socioeconomic disparities in life expectancy. When we assess how well single SES indicators approximate the estimates from our multi-factor approach, we find that the SES gradient is 50-150% steeper when combining multiple indicators rather than using income or education alone as normally done. Similarly, the share of variance in life expectancy explained by SES increases by 100-500%.

Combining just two indicators, for example income and education, substantially increases the share of variance explained, although it remains well below the level explained when all indicators are included. Among single indicators, wealth is the best 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, Lee & Weinstein 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 approach is distinct in two key respects: First and foremost, our focus is not on obtaining the best possible 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 multiple SES indicators, while deliberately excluding other types of information. Secondly, by extrapolating the mortality predictions to individual-level cohort life expectancy, we connect our analysis to the literature on inequality in life expectancy.

The following sections describe our data, method and results, and end with a conclusion in Section V.

II. 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 focus on individuals born between 1942 and 1944 and residing in Denmark at age 40 (years 1982-1984). 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 socioeconomic indicators at age 38, two years before we start analyzing mortality, to reduce the potential for reverse causality (Chetty et al. 2016; Kreiner, Nielsen & Serena 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), leaving 89,927 men and 101,466 women.[4] 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 factor used in the analysis (see the appendix for additional details). We focus on individual rather than household measures of SES, with results using household variables provided in the appendix.

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

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. 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 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. 2014). Our IQ measure is the Danish Armed Forces

[4] The occupation classification used includes categories for unemployment and retirement. Hence, the analysis sample is not restricted to individuals in employment at age 38.

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 10% 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.

III. Method

Conventional studies of life expectancy inequality compare groups defined by a single cardinal SES indicator, such as income ranks or years of education. In this paper, we are interested in capturing all relevant SES-related variations 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 longevity of those still alive by 78 is unobserved. To address this issue, we instead use machine learning to predict mortality in the ages we observe and then use survival models to extrapolate these estimates to cohort life expectancy.

We start by splitting the data into a training sample (60%) and a test sample (40%). We then construct a panel dataset for the training sample in which we follow individuals from age 40 to 78 or until death. Furthermore, in the training data set, we create a balanced sample with 50% mortality by randomly oversampling mortality cases. We do this because fitting an algorithm is easier when the two possible outcomes (dead, 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 is an ensemble method using decision trees as the base learner. In other words, the algorithm works by creating and averaging predictions across many decision trees. 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, Csörgö & Martínez-Muñoz 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).

We run the final XGBoost model on the balanced training data set using the optimal hyperparameter values.[5] By construction, the resulting algorithm predicts mortality in 50% of cases, exceeding actual mortality rates in the unbalanced datasets. To convert the predictions to match average mortality in the test sample, we use the unbalanced training dataset and run logistic

[5] 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.

regressions of mortality on a fifth-order polynomial of the XGBoost mortality predictions, interacted with age. The age interactions ensure that the predictions capture the age pattern of mortality.

We now have individual-level mortality predictions for each age from 40 to 78. Next, we run individual-specific Ordinary Least Squares (OLS) regressions of log predicted mortality on age (Gompertz 1825). From these regressions, we obtain individual-level estimates of the intercept and slope of predicted log mortality. To predict mortality beyond age 78, we plug these estimates into a Gompertz survival model.[6]

The Gompertz survival model is based on the proportional hazards model and assumes that the hazard rate $h(a)$ evolves exponentially with age a according to $h(a) = \lambda\, exp(\gamma a)$, where λ and γ are parameterized as $\lambda = exp(\alpha_0 + \alpha_1 x_1)$ and $\gamma = \beta_0 + \beta_1 x_2$.

After taking logs, we obtain:

$$\log(h(a)) = \alpha_0 + \alpha_1 x_1 + (\beta_0 + \beta_1 x_2) \cdot\ a$$

Hence, in this model, the intercept of log mortality depends on x_1, while the age slope depends on x_2. We use as x_1 and x_2, our OLS estimates of the intercepts and slopes of predicted log mortality.[7]

We run the Gompertz survival model on the training sample and use the coefficients to predict mortality, survival, and life expectancy in the test sample, which has not been used elsewhere in estimation. Therefore, we can use the test sample to assess the out-of-sample accuracy of our final models (see Appendix Figure 1).

---

[6] While we could predict life expectancy directly from the individual-level log mortality regressions, we use survival models to reduce the influence of individual-level noise.
[7] 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.

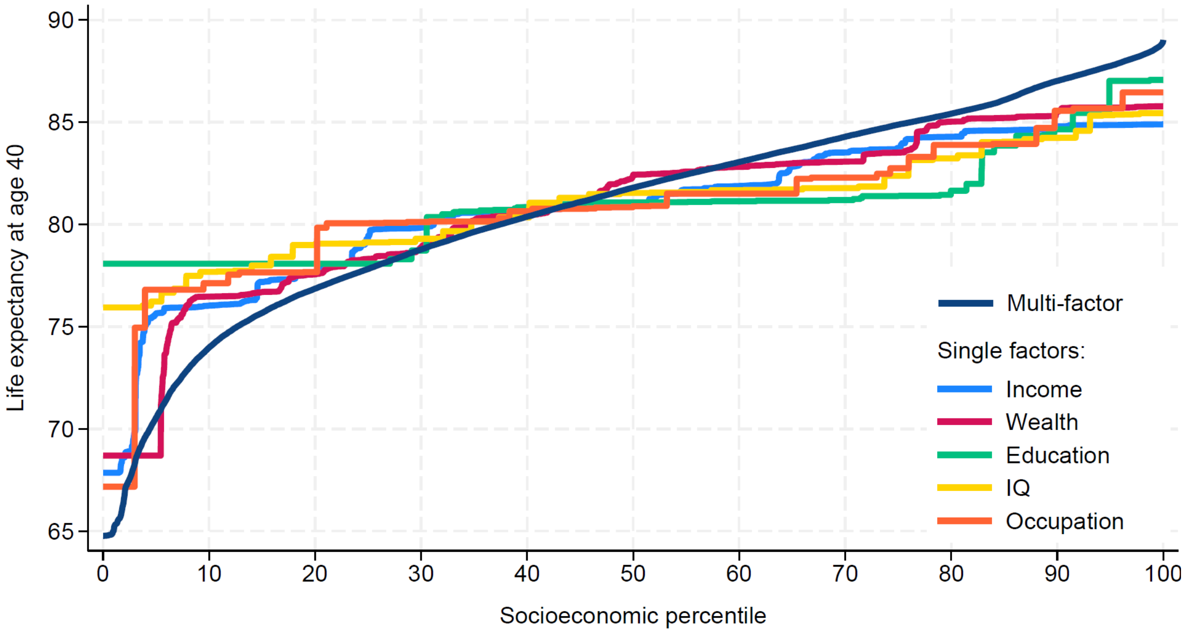
Note that our approach allows covariates (x_2) to change the slope of log mortality, in contrast to the proportional hazard model. More direct implementations of survival analysis in machine learning, such as DeepSurv (Katzman et al. 2018) and survival implementations in XGBoost and random forest rely 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).

IV. Results

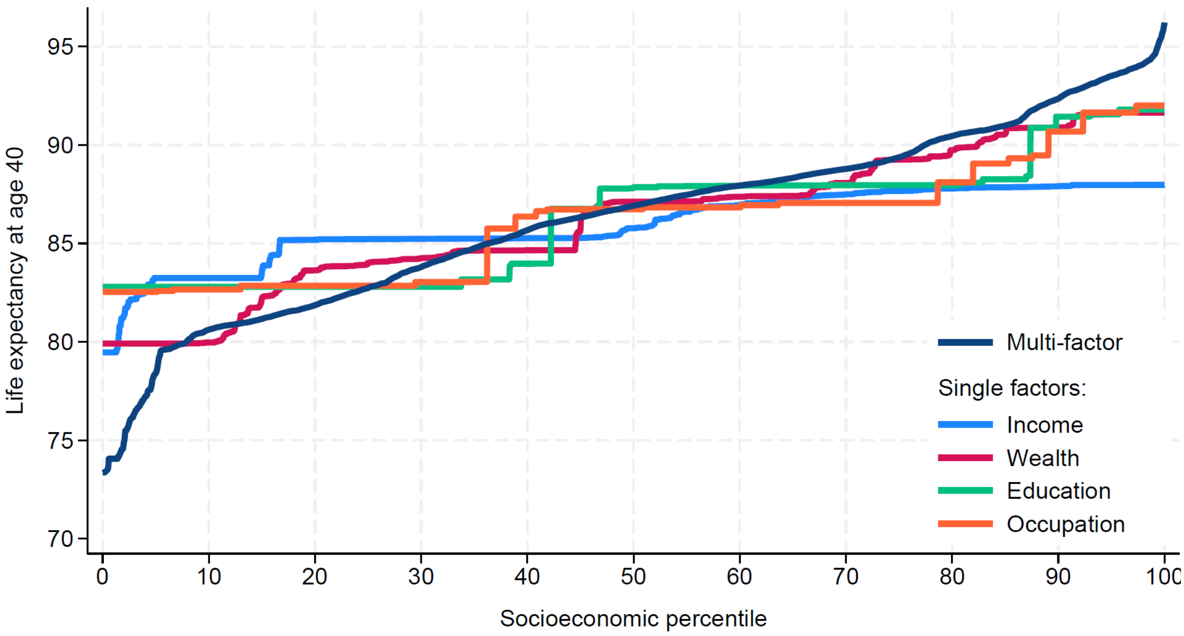
Figure 1 shows cohort life expectancy at age 40 by socioeconomic rank for men and women. The figure is constructed in the following way: For the line denoted Multi-factor, we first use all socioeconomic information to run the machine-learning approach described above. Using the estimated algorithm, we 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.

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 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 predictive power of the single-factor approach if there is a non-monotonic relationship between the SES indicator and life expectancy. This is most relevant for wealth (see Appendix Figure 2) where those with close to zero wealth live shorter lives than those with negative wealth, which includes the self-employed with large business losses. Therefore, the machine-learning algorithm assigns individuals with near-zero wealth with the lowest SES rank.

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 7 to 9 years. Interestingly, the estimated life expectancies based on education or occupation alone come closest to the multi-factor estimate in the top, while the estimates based on income or wealth come closest to the multi-factor estimate in the bottom.

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%.

Next, we assess the share of the variance in life expectancy attributed to SES, as a non-parametric summary measure of the importance of SES inequality. This is calculated as the ratio of the variance in average life expectancy between 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 suggests that while income approximates SES well at the extremes, wealth is better at capturing SES differences in the middle of the distribution.

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 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.

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	$75-87$	$78-85$	$76-85$	$76-85$	$78-84$	$77-86$
Linear gradient	0.172	0.078	0.116	0.135	0.085	0.107
SES share (%) of var.	11.8	2.6	5.8	7.8	2.7	5.4
Women						
Min-Max	$73-96$	$83-92$	$79-88$	$80-92$		$83-92$
P10-P90	$81-92$	$83-91$	$83-88$	$80-91$		$83-91$
Linear gradient	0.157	0.099	0.059	0.121		0.094
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 10^{th} and 90^{th} percentile. 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 between the between-SES-group variance in life expectancy to the overall variance in age at death.

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 the 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).

To understand why the multi-factor approach performs better, Table 2 examines who ends up with predicted life expectancies at the top and bottom. 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. However, around 80% rank in the top 10% if we instead compute an individual's average rank across all socioeconomic indicators. In other words, 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.

This pattern is even more visible 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 70% belong to the top 1% when ranked by the average of all indicators.

Table 2. Characteristics of individuals in the top and bottom of predicted life expectancy

	Men			Women		
	Bot. 10	Top 10	Top 1	Bot. 10	Top 10	Top 1
Income	54	38	4	38	23	3
Wealth	62	38	8	48	37	7
IQ	30	41	15			
Education	21	47	8	25	62	5
Occupation	44	40	13	20	46	4
Avg. rank	72	79	68	81	72	23
Avg. rank (ex. educ, IQ)	82			70		

Note: The table reports the percentage share of individuals in the bottom 10, top 10, and top 1 percent of predicted life expectancy (based on the multi-factor approach) who belong to the bottom 10, top 10, and top 1 percent, respectively, of the separate distributions of the various single SES indicators. In the "Average rank" row, individuals are ranked based on their average position across these separate SES indicators. Hence, individuals at the top have, on average, high ranks across all indicators. The "Average rank (ex. educ. and IQ)" row excludes education and IQ when calculating the average position across indicators.

We observe the same pattern 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. Among men at the bottom, it turns out that the average of income, wealth, and occupation is even more predictive than the average of all five indicators, as education and IQ are weak predictors of low life expectancy. The strength of the machine-learning approach is that it observes such patterns and derives the best prediction from the available information.

Combining a Limited Set of Socioeconomic Indicators

To examine the contribution of individual indicators and the extent to which using a subset of them can approximate the full estimate, we present two additional analyses in Table 3.

The upper pane of Table 3 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 contributes with little additional variation conditional on the other four indicators. While occupation alone accounts for 5.4% of the variation in life expectancy (according to Table 1), excluding it from the multi-factor analysis only decreases the share accounted for by the multi-factor approach from 11.8% to 11.4%. Income is somewhat more important, but the largest decrease in the share occurs when we remove wealth. In this case, the share accounted for by SES drops from 11.8% to 8.6%, demonstrating that wealth's contribution cannot be substituted by the four remaining indicators.

Table 3. 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	11.5	11.0	8.6	11.5	11.4
Women	7.1	8.1	6.4		8.2
2. *Combining two indicators*					
<u>Men</u>					
Education		7.2	9.8	4.0	6.8
Income			10.4	7.2	8.1
Wealth				9.8	10.6
IQ					6.5
<u>Women</u>					
Education		5.2	7.1		5.4
Income			6.0		4.2
Wealth					6.7

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 between 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 pane of Table 3 shows the share of the variance in life expectancy attributed to SES when we combine only two of the SES indicators. For example, for men, we account for 7.2% of the variance when combining education and income, compared to 2.6% and 5.8% when using education or income alone. Hence, the share is considerably higher when combining these two commonly used indicators. We saw previously that wealth is the most predictive single indicator, accounting for 7.8% of the variance. This percentage increases to 10.5% when we add income or occupation, illustrating that with 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.

For most two-factor combinations, the SES indicators partially substitute for one another, so the share of the variance accounted for by two indicators together is smaller than the sum of the shares accounted for by each individual indicator. A notable exception is the combination of education and income for women, which accounts for 5.2% of the variance, exceeding the sum of their individual contributions (3.7% and 1.3%). In this case, interaction effects increase the predictive power above the sum of each component.

Robustness

In this section, we describe robustness checks reported in the Appendix. To include information on IQ, we have focused our analysis on men who participated in mandatory military recruitment assessments (85% of the selected cohorts). This sample restriction may induce selection bias since participation is non-random, which further implies that sampling differs between men and women. Appendix Table 1 shows results for all men, including those without IQ test scores. The results are very similar. For example, the multi-factor estimates still start at a life expectancy of 65 years for those with the lowest SES and increase to 89 years for those with the highest SES. We extend this analysis by using household variables instead of individual-level variables, e.g., the average income of spouses (Appendix Table 2). Using household income increases the share of variance accounted for by SES for women, showing that household-level SES indicators can be more predictive than individual-level SES indicators. Still, the share accounted for by SES is significantly larger for the multi-factor approach than for the single-factor approach.

Following previous studies (Chetty et al. 2016; Kreiner, Nielsen & Serena 2018), we measure SES two years before starting to measure 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 3, 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 approach is very similar to Table 1.

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 4).

We have also repeated our main analysis where we use survival until age 78 as the outcome, which we observe in the data. Appendix Figure 3 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 33% at the lowest SES level to 78% at the highest, compared to a 42%-74% span across education groups and a 58%-76% span across income groups.

V. 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 gradient 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.

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.

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.

Chen, Tianqi, and Carlos Guestrin (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16)*, 785–794. ACM.

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.

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, 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.

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

Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger (2018). 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., Richards, S., & Cutler, D. M. (2008). The gap gets bigger: changes in mortality and life expectancy, by education, 1981-2000. *Health affairs (Project Hope)*, *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 Schirle, Tammy (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 (2023). *World Social Report 2023: Leaving No One Behind in an Ageing World.* New York: United Nations Department of Economic and Social Affairs.

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

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. 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., 2014). 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., 2014).

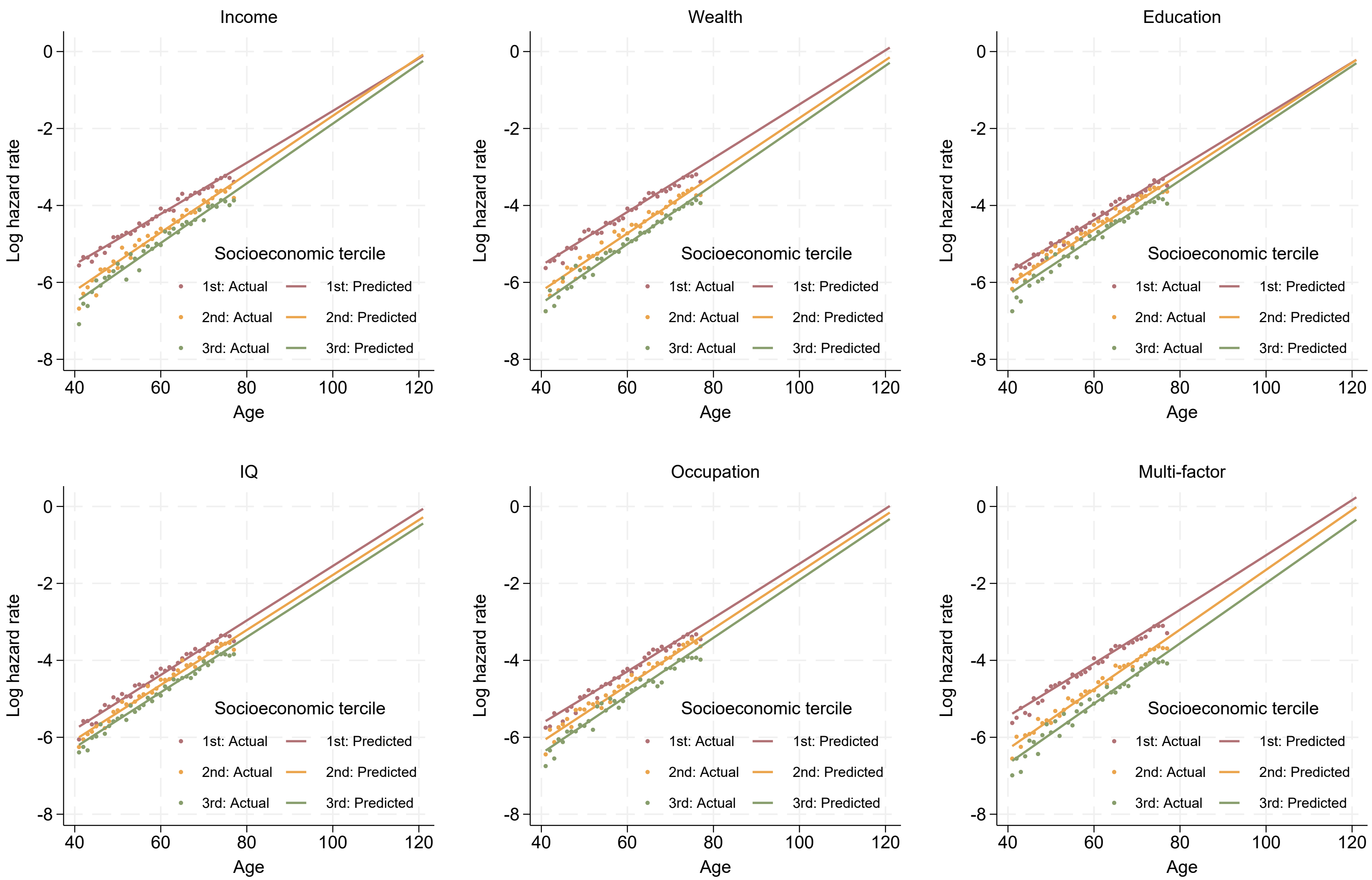
Occupation

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.

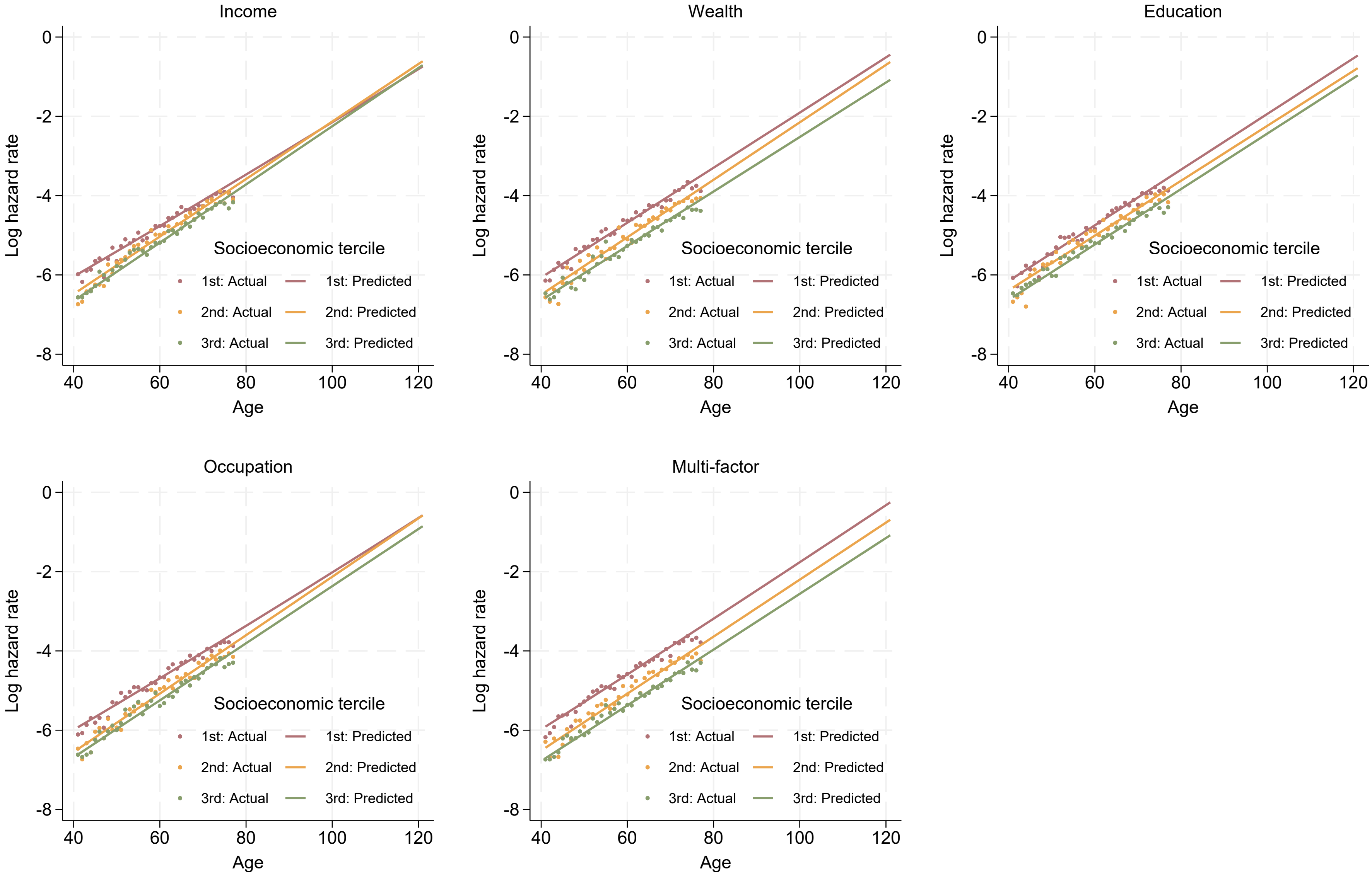
Appendix Figures and Tables

Figure A1: Gompertz approximations by age and SES for men



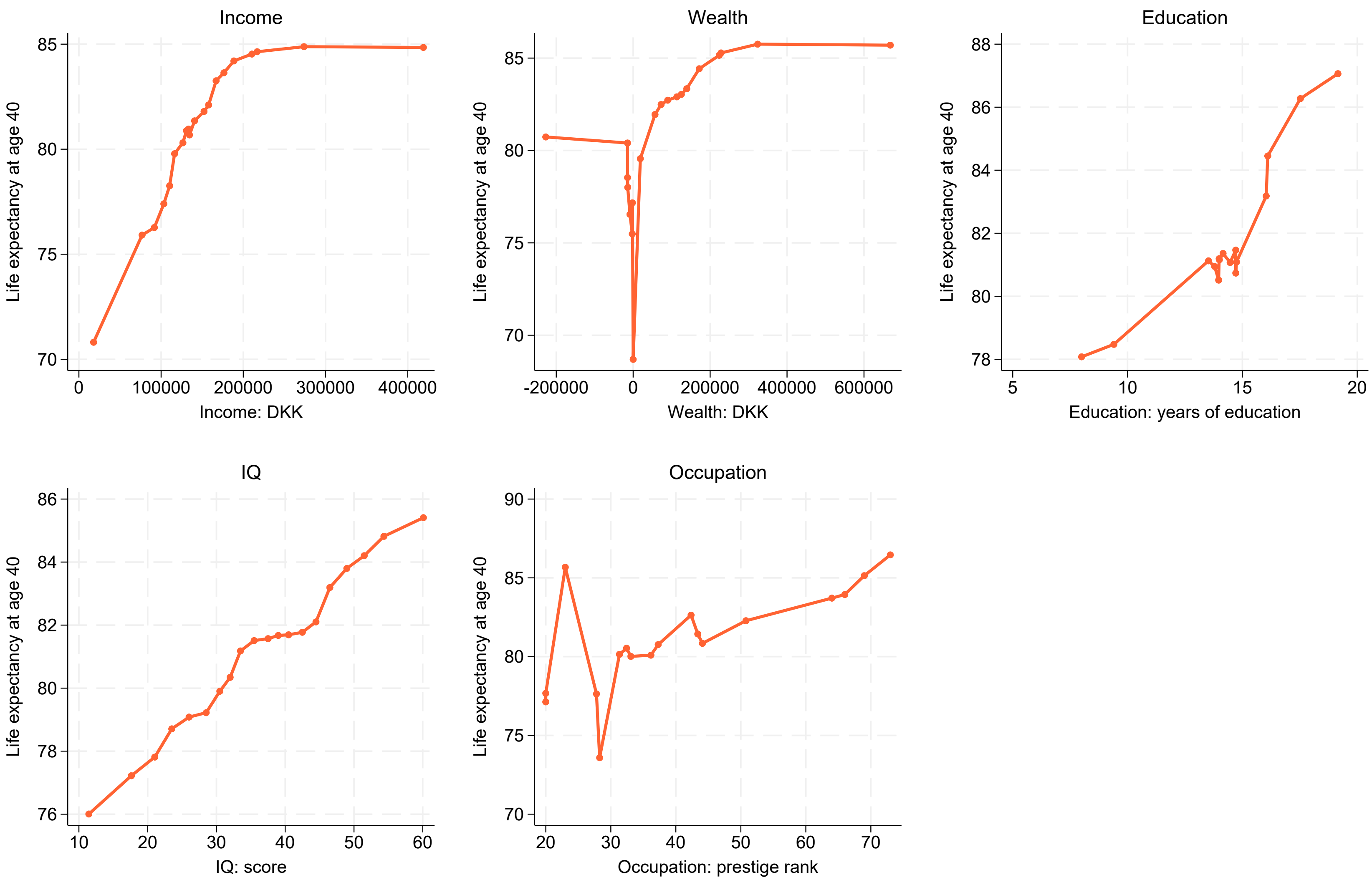
Note: The figure plots predicted and actual log mortality hazard rates by age and SES tercile. Terciles are constructed from predicted life expectancy based on the training data and the Gompertz model. The lines represent predicted hazard rates based on the training data. The dots represent Kaplan-Meier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

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



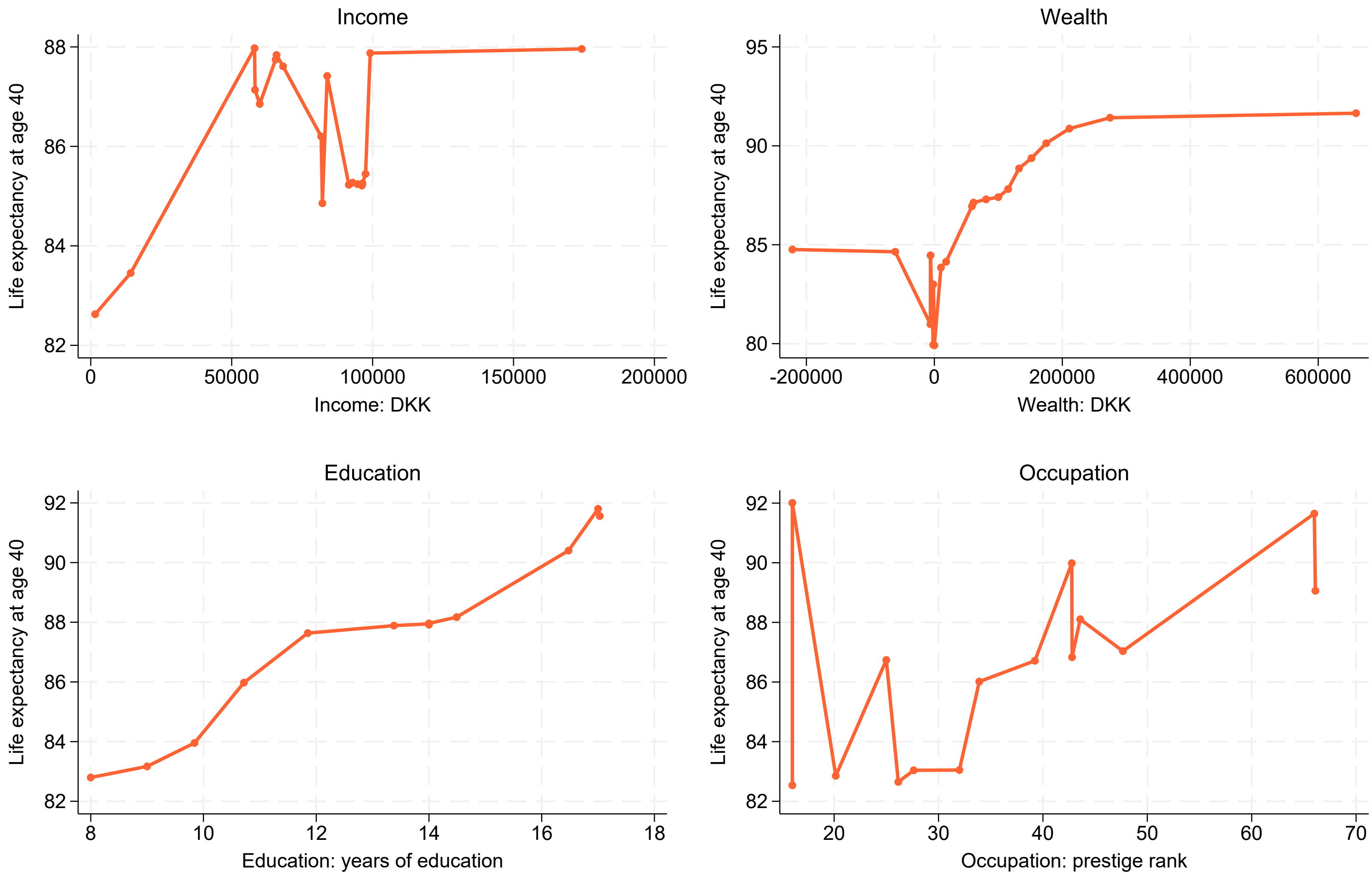
Note: The figure plots predicted and actual log mortality hazard rates by age and SES tercile. Terciles are constructed from predicted life expectancy based on the training data and the Gompertz model. The lines represent predicted hazard rates based on the training data. The dots represent Kaplan-Meier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

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



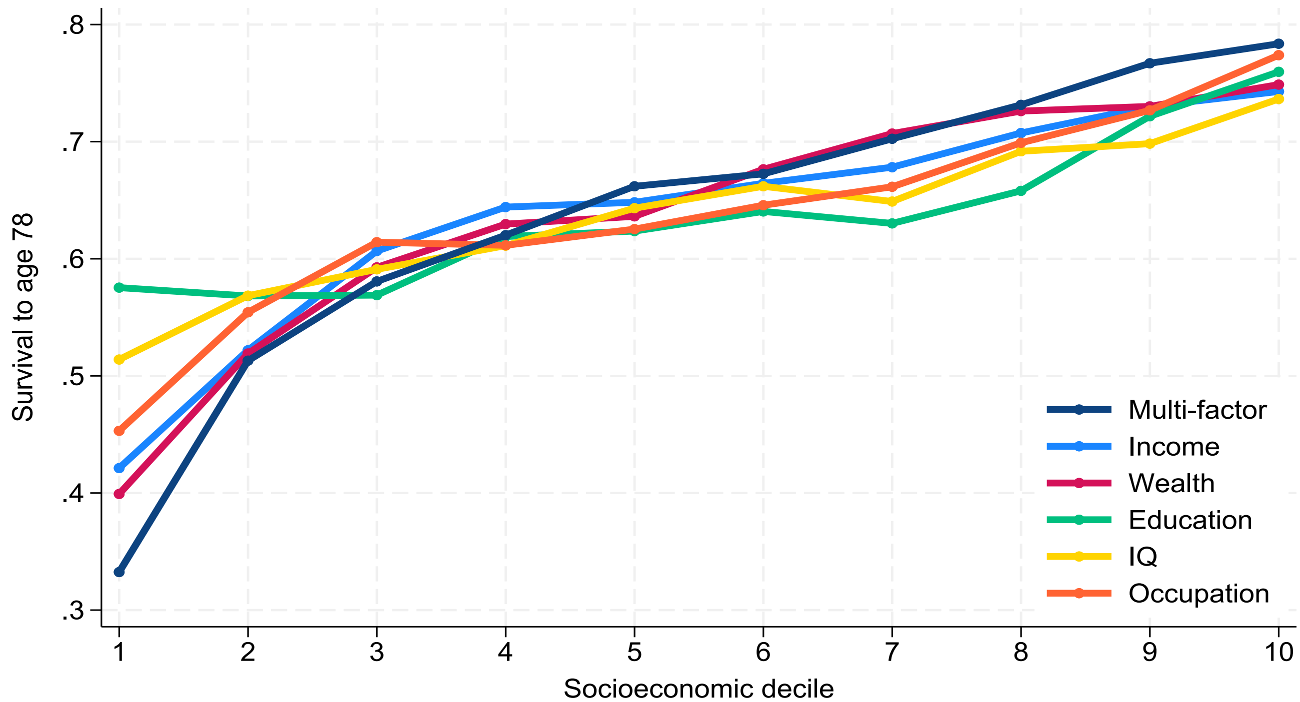
Note: The figure plots deciles of predicted life expectancy against the within-decile mean of the SES indicators used to make the prediction. The SES deciles are constructed from predicted life expectancy based on the training data and the Gompertz model.

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

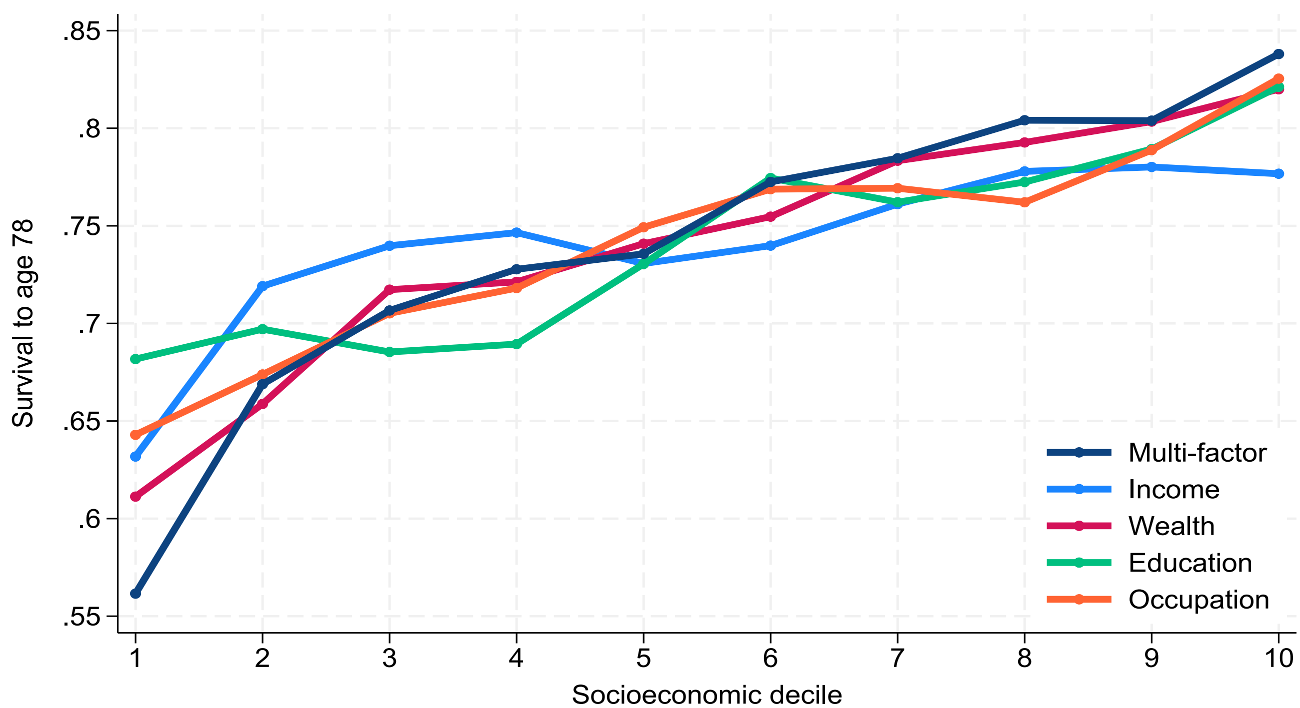


Note: The figure plots deciles of predicted life expectancy against the within-decile mean of the SES indicators used to make the prediction. The SES deciles are constructed from predicted life expectancy based on the training data and the Gompertz model.

Figure A3: 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: Life expectancy inequality for men without IQ restriction

	Multi-factor	Education	Income	Wealth	Occupation
Min-Max	65 - 89	77 - 87	67 - 84	69 - 86	67 - 86
P10-P90	73 - 86	77 - 85	75 - 84	76 - 85	77 - 85
Linear gradient	0.178	0.089	0.120	0.139	0.123
SES share (%) of var.	12.5	3.2	6.5	8.1	6.8

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 combination of all indicators.

Table A2: Life expectancy inequality, using household averages of socioeconomic indicators

	Multi-factor	Education	Income	Wealth	Occupation
Men					
Min-Max	65 - 90	76 - 87	67 - 84	69 - 86	68 - 89
P10-P90	73 - 86	76 - 83	79 - 84	76 - 85	76 - 85
Linear gradient	0.172	0.096	0.091	0.138	0.123
SES share (%) of var.	11.9	3.6	4.4	8.0	6.7
Women					
Min-Max	73 - 98	78 - 94	76 - 91	80 - 92	75 - 94
P10-P90	80 - 92	83 - 91	83 - 90	81 - 91	82 - 90
Linear gradient	0.167	0.122	0.085	0.124	0.118
SES share (%) of var.	9.3	5.0	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 A3: 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	77 - 86	69 - 87
P10-P90	74 - 87	78 - 85	76 - 85	76 - 86	78 - 84	77 - 85
Linear gradient	0.160	0.075	0.101	0.131	0.080	0.102
SES share (%) of var.	10.4	2.4	4.2	7.3	2.5	4.4
Women						
Min-Max	75 - 96	84 - 92	85 - 89	81 - 93		82 - 94
P10-P90	81 - 92	84 - 92	85 - 88	81 - 91		83 - 90
Linear gradient	0.145	0.097	0.041	0.121		0.093
SES share (%) of var.	7.4	3.6	0.6	5.1		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 A4: Life expectancy inequality, using Kannisto model for extrapolations

	Multi-factor	Education	Income	Wealth	IQ	Occupation
Men						
Min-Max	65 - 91	78 - 89	68 - 87	68 - 85	77 - 85	67 - 88
P10-P90	74 - 88	78 - 86	76 - 84	77 - 85	78 - 84	79 - 86
Linear gradient	0.182	0.078	0.101	0.129	0.080	0.097
SES share (%) of var.	12.8	2.9	4.7	7.6	2.5	5.2
Women						
Min-Max	76 - 99	83 - 92	80 - 89	80 - 91		81 - 95
P10-P90	80 - 95	83 - 91	84 - 88	80 - 91		81 - 90
Linear gradient	0.178	0.107	0.055	0.130		0.112
SES share (%) of var.	10.7	4.2	1.1	5.8		4.8

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, a:
Gompertz: $h(a) = \lambda e^{\gamma a}$. Kannisto: $h(a) = \frac{\lambda e^{\gamma a}}{1+\lambda e^{\gamma a}}$.