Heterogeneity in Expected Longevities[*]

Josep Pijoan-Mas

CEMFI *and* CEPR

José-Víctor Ríos-Rull

University of Minnesota
Federal Reserve Bank of Minneapolis
CAERP, CEPR, *and* NBER

July 2014

Abstract

We develop a new methodology to compute differences in the expected longevity of individuals of a given cohort who are in different socioeconomic groups at a certain age. We deal with the two main problems associated with the standard use of life expectancy: that people's socioeconomic characteristics change and that over time mortality has gone down. Our methodology uncovers a large amount of heterogeneity in expected longevities, yet much smaller than what arises from the naive application of life expectancy formulae. We decompose the longevity differentials into differences in health at age 50, differences in the evolution of health with age, and differences in mortality conditional on health. Remarkably, education, wealth, and income are health protecting but have very little impact on two-year mortality rates conditional on health. Married people and nonsmokers, however, benefit directly in their immediate mortality. Finally, we document an increasing time trend of the socioeconomic gradient of longevity in the period 1992-2008, and we predict an increase in the socioeconomic gradient of mortality rates for the coming years.

JEL classification: I14; I24; J12; J14

Keywords: Expected longevity; Mortality; Health; Socioeconomic gradient

[*] We are grateful for comments from Pierre-Carl Michaud, Pedro Mira, Carolyn Wilkins; from attendants to seminars at Penn, Carlos III, and a LAEF conference on Health and Macroeconomics; and from the editor and referees. Inés Berniell provided excellent research assistance. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis, or the Federal Reserve System.

1 Introduction

It is very well documented today that mortality rates are negatively related to socioeconomic status. In a seminal work, Kitagawa and Hauser (1973) showed that mortality rates in 1960 in the United States were inversely related to education and income. Since then, a large body of literature has emerged confirming the socioeconomic gradient of mortality rates, which is found in education and income but also in wealth, labor market occupation, or marital status.[1]

Much less is known, however, about how socioeconomic differences in mortality rates aggregate over the life cycle and create differences in the life duration of people of a given cohort. One approach to addressing this issue is to use differences in *period life expectancies*, which mechanically aggregate socioeconomic differences in the mortality rates of a given calendar year. For instance, Brown (2002) and Meara *et al.* (2008) compute differences in period life expectancies according to education level, and Lin *et al.* (2003) and Singh and Siahpush (2006) do so for other measures of socioeconomic status. However, the use of period life expectancies has two significant drawbacks. First, period life expectancy measures do not account for the possibility that socioeconomic characteristics of individuals evolve over time, i.e. membership in a given population subgroup may change over the life cycle, and hence so do the relevant mortality rates. This is typically the case for any measure of socioeconomic status, except for education. Second, period life expectancies typically do not account for the fact that mortality rates tend to decline over time, and that this may happen at different rates for people in different socioeconomic groups. The time-changing problem of mortality rates has been addressed by Lee and Carter (1992) and other related methods by estimating an age-specific time component of mortality rates that can be used for extrapolation. However, the time effects on mortality rates may not be independent from the socioeconomic status of individuals, and more importantly, exploiting time series variation ignores important current observable information that may have significant predictive power. To sum up, the static picture that emerges from differences in period life expectancy by different socioeconomic groups

[1] See, for instance, Montez *et al.* (2011) and references therein for recent findings of mortality differentials by education level. Deaton and Paxson (1994) document the negative relationship between mortality and family income, after controlling for education. Attanasio and Hoynes (2000) show the negative relationship between mortality and wealth. The Whitehall studies have uncovered important mortality differentials according to the employment grade among British civil servants; see, for instance, Marmot, Shipley and Rose (1984) and Marmot *et al.* (1991). For mortality rates and marital status see Hu and Goldman (1990) and references therein.

may not be a good proxy of actual differences in expected longevities within a cohort of individuals.

The first contribution of this paper is to develop and implement a new measure of the expected duration of life of a given cohort —or *cohort life expectancy*— that addresses these two problems. We call this measure *expected longevity*, and document its socioeconomic gradient. In a first step, we exploit the panel structure of the Health and Retirement Study (HRS) to estimate age-specific survival rates conditional on a socioeconomic characteristic of interest z, as well as its age-specific transition probabilities (Section 4). The transitions for the socioeconomic characteristic z allow us to address the changes in socioeconomic status over the life cycle. The socioeconomic characteristics studied here include education, wealth, non-financial income, labor market status, marital status, and smoking. We then link the estimates for all different cohorts in the HRS in order to build expected longevities conditional on the characteristic z at age 50 for individuals born between 1941 and 1954 (who were of 50 years of age between 1992 and 2004). Our results uncover a large amount of heterogeneity in expected longevities across different measures of socioeconomic status, with the socioeconomic gradient being always steeper for men than for women. That said, these socioeconomic gradients are two to three times smaller than those that result from using period life expectancies. This confirms the importance of keeping track of the life-cycle evolution of individual socioeconomic characteristics in order to predict the life duration of a given population sub-group.

In a second step, we estimate age-specific survival rates conditional on a socioeconomic characteristic z and on individual (self-assessed) health h and age-specific joint transition probabilities for z and h (Section 5). The use of information on health h allows us to partly address the changes over time in survival rates and in the transitions for z and to take into account that the transitions on z may have a direct dependence on health. This is quantitatively important: the socioeconomic gradients in expected longevities that we find are larger (between 20% and almost 100%) than when we ignore information on health, and hence assume that mortality rates and transition functions are constant over time. The reason for this is that the gap in health condition at age 50 between the most and the least advantaged types has grown over time. Hence, our results point to an increase in the socioeconomic gradient of mortality rates at old ages for the coming years.

The second contribution of this paper, intimately linked to the first one, is to decompose the socioeconomic differences in health already present at age 50, changes in health that developed after age 50, and differences in two-year mortality rates unrelated to measured health. We find that around one-third of the expected longevity differentials for education and wealth categories are already there in terms of health differences at age 50, while the remainder is due to the health protection effect of education and wealth over the following years. Interestingly, the effects of education, wealth, and income on two-year mortality rates are very small or null after controlling for self-assessed health. This finding is surprising, as higher wealth, income, and education suggest a greater ability to pay for medical treatment. While financial resources may be health protecting, they do not appear to lower mortality at the onset of terminal diseases, nor do they appear to reduce death inducing accidents. In contrast, being married and a non-smoker significantly reduces mortality rates even after controlling for differences in measured health. This raises the question of what is exactly behind the survival advantage of married people.

Finally, our third contribution is to exploit the relatively long time span of the HRS to examine the time evolution of the socioeconomic gradient of expected longevity. We find relatively large increases in this gradient, although the precision of our estimates is low.

2 HRS data

The Health and Retirement Study is a biennial panel of individual level data, ranging from 1992 to 2010. The first wave of interviews was made to respondents born between 1931 and 1941, and their spouses regardless of age. New cohorts have been introduced over the years, both younger and older. The original cohort is named HRS. The new cohorts are the AHEAD (introduced in 1993, people born before 1924), CODA (introduced in 1998, people born between 1924 and 1930), WB (introduced in 1998, people born between 1942 and 1947), and EBB (introduced in 2004, people born between 1948 and 1953). The overall sample contains respondents aged 51 or older, plus spouses of any age.

Individuals of age 50 in the HRS data set are born in 3 different cohorts: 1941-1942 (HRS cohort, observed in 1992), 1947-1948 (WB cohort, observed in 1998), and 1953-1954 (EBB cohort, observed in 2004). Figure 1 provides a complete description of the age-cohort structure of the HRS target population. The length of the arrow reflects the maximum age range in which we observe the individuals of a given year of birth. The dashed diagonal lines indicate the year of observation. The patent unbalanced entry of new cohorts generates an age structure of the target population that is very different in every sample year, which prevents a clear linking of our findings to a specific year.

[Figure 1 about here.]

2.1 Sample selection

Our sample excludes individuals for which we cannot obtain race, sex, or education, and individual-year observations for which we cannot obtain self-rated health or the survival status into the next interview. We keep individual-year observations with a positive sampling weight —which represent the civilian non-institutionalised US population— and age range 50 to 94. We create separate samples for males and females because the slope of the socioeconomic gradient has been shown to be different among genders, see for instance Elo and Preston (1996) or Rogers *et al.* (2010). Our study focuses on white individuals due to the much smaller sample sizes for other race/ethnic groups. Socioeconomic gradients of health exist in all groups, but they are of different magnitude, see Crimmins *et al.* (2004) and references therein.

We estimate survival rates and transition functions conditional on education, marital status, labor market status, wealth, income, smoking status and self-rated health. Some of these variables present missing data for a few observations. We do not drop observations with missing data other than education and self-rated health, so sample sizes in different estimations may differ slightly. We drop the observations with missing data for self-rated health because we want the samples for Sections 4 and 5 to be identical. Overall, we have 53,362 individual-year observations for males and 67,453 for females, which correspond to 9,542 males and 11,236 females. Our sample period is 1992 to 2008 since no transition of any type can be observed in 2010. In Appendix A we describe the variable definitions and give more details about the sample selection.

3 Period life expectancies

The measurement of expected years of life for a subgroup of individuals of a given cohort is not an easy task. The crudest way of doing so is by aggregating the age-specific mortality rates—also known as life tables—of the population subgroup into *period life expectancies*. The period life expectancy at age 50 measures the average age of death for a hypothetical group of 50-year-olds, born at the same time and subject throughout their lifetime to the age-specific death rates of a particular time period, usually a given calendar year. The National Vital Statistical System (NVSS) computes the life tables for the U.S. population and reports period life expectancies for gender-race subgroups. The period life expectancy differs from expected longevity, in that the latter accounts for possible changes in a person's type and mortality rates over time. As such, the period life expectancy calculations provide a useful benchmark from which to assess the importance of these possible changes on the social gradient, with and without controls on self-reported health.

Due to the relatively small sample size of the HRS, we cannot use individuals born in different years to compute the death rate at every different age. Instead, we pool the data of all years to compute period life expectancies, using data from individuals in several cohorts to compute each age-specific survival probability. Since the HRS sample period is between 1992 and 2008, we view our period life expectancies as a weighted average of those reported by the NVSS between these years.

3.1 Average life expectancy

We start by computing the life expectancy at age 50 using pooled data from the HRS to estimate separate age-specific two-year survival probabilities γ_a for white males and white females. These probabilities are estimated with a logit model that includes a linear term in age. Details on this and all the remaining estimations can be found in Appendix B.1 and Appendix B.2. Let x_a be the number of people alive at age a out of a given initial population at age 50. Then, the life expectancy at age 50, e_{50}, can be computed as follows:

$$\begin{aligned} e_{50} &= \sum_{a \in A} \left[a \left(1 - \gamma_a\right) x_a \right] + 1 \\ x_{a+2} &= \gamma_a \, x_a \quad \forall a \geq 50 \\ x_{50} &= 1. \end{aligned}$$

Since the HRS is a biennial panel, all of our estimates refer to two-year periods. Due to the scarcity of data for very old individuals, we restrict our estimates to people up to age 94. Hence, we define $A \equiv \{50, 52, ..., 94\}$. Note that our formula for life expectancy is unconventional. In actuarial sciences life expectancy is typically defined as $e_{50} =$

$\sum_{a \in A} a \; {}_a\gamma_{50} + 1$, where ${}_a\gamma_{50}$ is the probability of survival to age a for an individual of age 50 and is computed by use of the conditional survival probabilities γ_a. We prefer to keep our formulation to preserve comparability with the formulas of expected longevity used in the later sections of the paper.

[Table 1 about here.]

In the first column of Table 1 we report the point estimate of the average life expectancy and its standard error.[2] We find that life expectancy at age 50 is 78.8 years for white males and 82.9 years for white females. These numbers square well with the life expectancies computed with the life tables reported by the NVSS in the years 1992 through 2008. In particular, the NVSS life expectancies between 1992 and 2008 range from 77.0 to 79.3 for white males and from 81.7 to 82.9 for white females.[3]

3.2 The socioeconomic gradient in life expectancy

The age-specific mortality rates vary substantially with variables related to socioeconomic status. As we discussed in the Introduction, it is well known that the more educated, the wealth-rich, the income-rich, and the married have lower mortality rates. It has also been shown that being active in the labor market is related to lower mortality rates, see for instance Lin *et al.* (2003). We can aggregate these differences in mortality rates by computing life expectancies for each group. In particular, we compute life expectancies conditional on a characteristic $z \in Z \equiv \{z_1, z_2, \ldots, z_n\}$ and obtain the difference $e_{50}(z_1) - e_{50}(z_n)$, where z_1 and the z_n are the most and the least advantaged types. We consider different sets Z of socioeconomic characteristics: education (college graduates, high school graduates, less than high school degree), wealth (quintiles of the distribution of total household net worth per adult), non-financial income (quintiles of the distribution of labor income plus employer pensions plus all government and social security transfers), labor market status (three categories: attached, which we define as working full time or being unemployed searching for a job; semi-attached, which we define as working part

[2] The standard errors are obtained by drawing 25,000 samples of parameter values from the estimated asymptotic distribution of the model parameters, and computing a life expectancy with each of them. See Appendix B.3 for details.

[3] The life expectancies we compute in the HRS should be somewhat larger than the ones reported by the NVSS because the HRS refers to the non-institutionalized population. For instance, Brown *et al.* (2012) find that life expectancies at 65 in the HRS are about one year larger than in the NVSS.

time or being semi-retired; and inactive, which includes inactive, retired and disabled individuals), and marital status (married and its complement).[4] In addition, we also consider smoking behavior, which is not a socioeconomic characteristic but a risky behaviour. However, because it is a habit, smoking tends to be very persistent over time and therefore longevity predictions conditional on smoking status can be computed with the same methodology that we develop here. Finally, to illustrate the versatility of our methods, we also look at a four-category variable created by combining marital status and smoking. The interpretation of all these life expectancies is the expected age of death of a hypothetical group of 50-year-olds with some characteristic $z = z_j$ who are subject throughout their lifetime to the z_j age-specific death rates of the current population alive.

To compute z-specific life expectancies, we first estimate age-specific two-year survival probabilities $\gamma_a(z)$ for every $z \in Z$. This involves estimating logistic regressions of survival against age, dummies for each $z \in Z$, and interaction terms between age and z in order to allow for the fall in survival due to age being different for different types z. The life expectancy $e_{50}(z_j)$ at age 50 for individuals whose z was equal to z_j at age 50 is then given by

$$\begin{aligned} e_{50}(z_j) &= \sum_{a \in A} \left[a \left[1 - \gamma_a(z_j)\right] x_a \right] + 1 \\ x_{a+2} &= \gamma_a \, x_a \quad \forall a \geq 50 \\ x_{50} &= 1. \end{aligned}$$

In Table 1, columns 2-7, we report the life expectancy differentials at age 50 together with the standard errors. We find huge socioeconomic gradients of life expectancies, all of them steeper for males than for females. At age 50, the difference in life expectancy between college graduates and individuals without a high school degree is 6.3 years for males (and 5.8 years for females); the difference between individuals at the top and bottom quintiles of the wealth distribution is 10.7 years (8.5); the difference between individuals at the top and bottom quintiles of the non-financial income distribution is 6.1 years (3.6); the difference between individuals strongly attached to the labor force and inactive individuals is 9.4 for males and 5.0 for females; the difference between married and nonmarried individuals is 4.6 years (2.4); the difference between nonsmokers and smokers is 7.2 years (6.2). Combining marital status and smoking also shows large differentials: the difference between married nonsmokers and nonmarried smokers is 10.2 years (7.5). These

[4] See Appendix A for the exact definition of all these variables.

results should not necessarily be interpreted in terms of causality, given that selection of healthier individuals into better socioeconomic groups is known to matter. The work in Section 5 uses a joint transition function of health and various other characteristics (like marital status), which addresses this issue.

4 Expected longevities

Period life expectancy for individuals with a certain socioeconomic characteristic z can be a biased measure of the length of life of those individuals when the characteristic of interest z changes over time. For instance, the period life expectancy calculation for the top income quintile is likely biased upwards. This is because it does not take into account the fact that some individuals will fall ill, drop out of the top income quintile, and then die (probably earlier than those who remain in the top income quintile). This is true even if the age-dependent mortality rates by type z do not change over time. This is not a problem with education, which is fixed at age 50. But wealth, income, marital status, and smoking behavior change substantially over the years.[5]

Accordingly, we start developing a measure of expected longevity at age 50 conditional on a given characteristic $z \in Z$ at age 50 that allows for changes in the characteristic z over the life cycle. In Section 5 we further extend this concept using information about health to account for the changing composition of the health of the population whose expected longevity we want to measure. Two elements are required for these calculations: age-dependent survival rates conditional on z, $\gamma_a(z)$, and age-dependent transition probabilities for state z, $p_a(z'|z)$. Of course, the latter is not needed for education. Let $x_a(z)$ be the fraction of people who are alive and of type z at age a out of a given population at age 50. Expected longevity $\ell_{50}(z_j)$ at age 50 conditional on $z = z_j$ at age 50 is then computed as

$$\begin{aligned}
\ell_{50}(z_j) &= \sum_{a \in A} \Big[a \sum_{z \in Z} [1 - \gamma_a(z)]\, x_a(z) \Big] + 1 \\
x_{a+2}(z') &= \sum_{z \in Z} p_a(z'|z) \;\; \gamma_a(z) \;\; x_a(z) \qquad \forall z' \in Z, \; \forall a \geq 50 \\
x_{50}(z_j) &= 1 \;\text{ and }\; x_{50}(z) = 0 \;\; \forall z \neq z_j.
\end{aligned}$$

[5] To point to particular example, 47% of white males with wealth in the top quintile of the distribution at age 50 were in the same quintile by age 65 (with most of the movers going to the second quintile). Or 88% of white males that were married at age 50 were also married by age 65. Even more important are the changes in labor market status because people clearly drop from the labor force as they age.

We use the same estimates of $\gamma_a(z)$ as in the previous section, and we estimate multivariate logistic regressions with the same regressors in order to compute the transition matrices $p_a(z'|z)$. The interpretation of these expected longevities is the expected age of death of a hypothetical group of 50-year-olds with some characteristic $z = z_j$ who are subject throughout their lifetime to the age-specific type-z transition rates and age-specific survival rates conditional on z of the current population alive.

[Table 2 about here.]

4.1 The socioeconomic gradient in expected longevity

In rows (1) of Table 2, we report the differentials in expected longevity for the same socioeconomic categories as in Table 1, except for education because education does not change after age 50. We find a very important amount of heterogeneity in expected longevities at age 50, with all differences statistically different from zero. The largest differences in expected longevity are between married men who are non-smokers and non-married men who are smokers (4.6 years), and between men in the top and bottom quintiles of the wealth distribution (3.6 years). The socioeconomic gradient is generally more important for men than for women, except for non-financial income which is relatively small at 0.8 years for both.

The degree in heterogeneity based on socioeconomic characteristics is substantially less when changes in these characteristics after age 50 are accounted for, as shown by the large disparity in results between Tables 1 and 2. In fact, the estimates assuming static socioeconomic characteristics (period life expectancy) are between 2 and 7 times higher for men and 2 and 6 times higher for women than the estimates that account for changes in socioeconomic characteristics (expected longevity). Since life expectancy is a measure of expected longevity that imposes an identity matrix for the transition $p_a(z'|z)$ of characteristic z, the lower the mobility between groups, the smaller the difference between the gradients in life expectancy and expected longevity. Our findings using the HRS data indicate how huge is the empirical importance of accounting for mobility between groups.

4.2 The socioeconomic gradient within education groups

More educated people tend to have more favorable socioeconomic characteristics (e.g., be richer and earn higher income, divorce less). In order to get a better sense of the extent to which the differences in expected longevity associated with different socioeconomic variables are different from each other, we also compute our measures of expected longevity by education group. In rows (2) and (3) of Table 2 we report these differentials for college graduates and for individuals without a high school degree. We find that the differentials remain large within education groups, still more so for men than for women. In all cases, the differentials are larger within individuals without a high school degree than within college graduates.

Finally, the longevity differentials between college graduates with the most advantaged type z and individuals without a high school degree with the least advantaged type are reported in row (4) of Table 2. The differentials are larger than the average differential by education group, underscoring the importance of characteristic z beyond education in estimating expected longevity.

4.3 Time trends

The data from NVSS show an upward trend in life expectancies, with life expectancies at age 50 having increased by 2.2 years for white males and by 1.0 years for white females between 1992 and 2008. There is no reason to expect that all socioeconomic groups have shared equally in this improvement. Indeed, an increase in the educational gradient of mortality and life expectancy has been widely documented for the US over these years, see for instance Preston and Elo (1995), Meara *et al.* (2008), or Montez *et al.* (2011). In order to uncover possible time changes in the socioeconomic gradient of expected longevities, we add the calendar year to our estimates of the age-specific survival rates and the age-specific transition probabilities for types z. By adding a linear year term independent of age but z-type dependent, we allow for both survival probabilities and mobility between types to change over time, and to do so differently for different types. Instead, we restrict the time changes in survival and mobility to be homogeneous across ages.[6] With these

[6] Allowing for interactions between age, type, and year would increase the parameterization of our logit and multilogit models beyond tractability. In addition, the rationale for interacting time effects with age comes from the evidence that long-run gains in survival rates are different at different ages. However, these findings relate to both age differences and time intervals much wider than ours. See Lee and Carter

estimates, we can compute expected longevities as in Table 2 but specific to every year in our sample. The expected longevities of a particular year are consistent with individuals facing throughout their remaining life the mortality rates and the transition matrices of the given year.

In Table 3 we report the average life expectancy and the socioeconomic gradient in expected longevity for the first and last years of our sample period, as well as their changes over these 18 years. Our estimates show an increase of 2.1 years in the life expectancy at age 50 for white males between 1992 and 2008, with a standard error of 0.7. The corresponding increase reported by the NVSS is 2.2 years. Hence, our HRS sample captures the population trend for white males very well. By contrast, the life expectancy for white females falls by 0.3 years in our sample, whereas it increases by 1.0 years in the NVSS data.[7]

[Table 3 about here.]

The estimated longevity differentials have all been increasing over time with the exception of non-financial income. In particular, the educational differential increased by 1.7 years for white males, and 2.4 for white females, although the standard errors are of the same order of magnitude. This is consistent with Preston and Elo (1995), which shows evidence of an increase in the education gradient of mortality rates between 1960 and the early 1980s.[8] Montez *et al.* (2011), using the National Health Interview Survey Linked Mortality File, also find an increase in the mortality gradient for both white men and white women during the period 1986-2006. Meara *et al.* (2008) report that the education differential of life expectancy at age 25 increased by 0.9 years for white males and 1.1 years for white females between the 1980's and the 1990's, and by 1.6 years and 1.9 between 1990 and 2000.[9] This consistency in evidence supports our view about the value

(1992) for details.

[7] Despite the fact that the standard error associated to the change in female life expectancy is large (0.6 years), this discrepancy between the HRS and the NVSS is worrisome. In a sense, we are stretching the HRS to its limits. As shown in Figure 1, information of deaths for old individuals contain limited time variation. For instance, individuals aged 85+ come only from the original AHEAD cohort. This problem is more acute for women, who on average die 4 years later.

[8] Preston and Elo (1995) showed that the education gradient of mortality rates computed with the NLMS between 1979 and 1985 is larger than the one obtained by Kitagawa and Hauser (1973) with the death certificates and census data of 1960.

[9] The results for the 1980s and 1990s are based on data from the NLMS, whereas the comparison between 1990 and 2000 is based on data from the death certificates in the Multiple Cause of Death files.

of looking at time trends with the HRS, despite the lack of precision of estimates using this data.

Our results for the other measures of socioeconomic status are novel, and hence they paint a wider picture. During the decades of the 1990s and 2000s the college premium in the labor market and income inequality have increased —see Heathcote *et al.* (2010)—, and so has wealth inequality — see Díaz-Giménez *et al.* (1997) and Díaz-Giménez *et al.* (2011). Hence, a tempting conclusion is that the increase in income and wealth inequality is behind the increase in the socioeconomic gradient of expected longevity. However, our results show that the gradients for marital status and smoking have also increased over this period. This might be due to the correlation between marital status or smoking with income-related variables. But it may also be due to an increase in the selection of long-lived individuals into marriage and nonsmoking.

4.4 Summary

Our findings of Section 4 can be summarised as follows. First, socioeconomic differences in period life expectancies are very poor predictors of differences in expected longevities when the socioeconomic characteristics of interest change over the life cycle. Second, while education is the most important variable to predict longevity differentials, other socioeconomic variables such as wealth, labor market status, or marital status carry independent relevant information. Third, there is more heterogeneity within males than within females and within less educated than within more educated individuals. And fourth, these differentials in expected longevity between socioeonomic groups have been increasing substantially during the last 17 years.

5 Expected longevities with information about health

We use self-assessed health in order to improve our measures of expected longevity. Self-assessed health has been found to be a very important determinant of survival probabilities even after controlling for socioeconomic characteristics and measured health conditions, see for instance Idler and Benyamini (1997) and Idler and Benyamini (1999). In addition, self-rated health is a very interesting measure of health because it is also present in several other data sets of individual survey data commonly used by social scientists, such as the

English Longitudinal Study of Ageing (ELSA).[10]

The HRS asks respondents to evaluate their general health level among five categories: excellent, very good, good, fair and poor. We use these data to estimate age-dependent health h and type z survival rates, and age-dependent joint health h and type z transition functions. Due to sample size restrictions, we will use our pooled data of all cohorts to compute these objects. However, we will compute expected longevities at age 50 only for those cohorts whose health distribution is observed at age 50, that is to say, for the cohorts 1941-42, 1947-48, and 1953-54. Hence, these expected longevities assume that, conditional on health, the future survival rates by type —and their transition functions— of the 1941-1954 cohorts will be as the ones observed today for older cohorts. But since the underlying health distribution may differ, the actual predicted survival rates —and the transition functions— will also do so.

5.1 Average expected longevity

We start by computing the average expected longevity ℓ_{50}^h using information on health and without conditioning on any type variable z. To do so, we estimate age-dependent survival as a function of health $\gamma_a(h)$, an age-dependent health transition function $p_a(h'|h)$, and the initial health distribution $\varphi_{50}(h)$. Logistic and multinomial logistic regressions are estimated for the survival and transition functions, using as regressors a linear term in age, dummies for each $h \in H$, and interaction terms between age and h. We then compute the expected longevity ℓ_{50}^h as

$$\begin{aligned}
\ell_{50}^h &= \sum_{a\in A}\left[a\sum_{h\in H}\left[1-\gamma_a(h)\right]x_a(h)\right]+1, \\
x_{a+2}(h') &= \sum_{h\in H} p_a\left(h'|h\right)\ \gamma_a(h)\ x_a(h), \qquad \forall h' \in H,\ \forall a \geq 50, \\
x_{50}(h) &= \varphi_{50}\left(h\right).
\end{aligned}$$

This new measure of expected longevity, ℓ_{50}^h, is the expected remaining life of a given cohort of individuals that face the same age-dependent mortality rates conditional on health $\gamma_a(h)$ and the same age-dependent evolution of health $p_a(h'|h)$ as the current old,

[10] Others include the Medical Expenditure Survey (MEPS), the National Health Interview Survey (NHIS), the National Health and Nutrition Examination Survey (NHANES), the National Longitudinal Study of Youth (NLSY), the Survey of Health Ageing and Retirement in Europe (SHARE), or the Panel Study of Income Dynamics (PSID).

but may differ on the initial distribution of health $\varphi_{50}(h)$. Compared with the *period* life expectancy e_{50} of Section 3, the measure ℓ_{50}^h takes into account the possibility that the 50-year-olds born between 1941 and 1954 may face in the future mortality rates γ_a that are different from those faced by the current old, who were born earlier. The measure ℓ_{50}^h does so through the observed differences in health status, instead of relying on extrapolation from time series regressions as in the Lee and Carter (1992) type of methods. While this approach does not attempt to extrapolate $\gamma_a(h)$ and $p_a(h'|h)$ to future dates, it still incorporates some of the improvements in survival over time: using our method, if the health distribution at age 50 of the 1941-54 cohorts is better than it was for the older ones when they were of age 50, it must be the case that $\ell_{50}^h > e_{50}$.

In the first column of Table 4, we report the expected longevity ℓ_{50}^h. We find the expected longevity at age 50 to be 78.8 years for white males and 83.0 years for white females. These values are nearly identical to the life expectancies e_{50} computed in Section 3 (see Table 1, column 1), which were 78.8 and 82.9, respectively. This indicates the absence of relevant differences in initial health at age 50 in favor of the 1941-53 cohorts. If the large trend gains in life expectancy over the last years were to extend into the future, it would therefore not be through the better health of the 50-year-olds, but rather through improvements over time of the age-dependant survival function $(\gamma_a(h))$ and the age-dependent transition function $(p_a(h'|h))$.

[Table 4 about here.]

5.2 The socioeconomic gradient in expected longevity

To use information on health to improve on our measures of the socioeconomic gradient of expected longevity computed in Section 4.1 we will need three different objects: a) the health distribution at age 50 for every type z, $\varphi_{50}(h|z)$; b) the age-dependent joint health and characteristic z transition matrix, $p_a(z', h'|z, h)$; and c) the age-dependent survival rates conditional on health and characteristic z, $\gamma_a(z, h)$. We use the same logit and multilogit models as in the previous section, with dummies for each element in Z and H, and their interaction with age.[11] For the transitions, the outcome variable is given by

[11] Some authors choose to estimate the health and survival functions together through an ordered logit, thinking of death as an extra (and absorbing) health state; see, for instance, Yogo (2009). Our specifications has two advantages. First, it is designed to estimate not only the effects of the type variables z into health, but also the evolution of the type variables z and how this is affected by health itself. Second,

all the elements in the set $Z \times H$, so our estimates allow for health changes to have a causal impact on socioeconomic characteristics. This is important. The potential impact of health on wealth, income, and labor market status has been largely documented, see for instance the surveys by Smith (1999) and Currie and Madrian (1999). In addition, Hu and Goldman (1990) provide evidence of the importance of selection of less healthy individuals into single and divorced status. Let $x_a(z,h)$ be the fraction of people who are alive and of type z with health h at age a out of a given population at age 50. Given these objects, we can build expected longevity $\ell^h_{50}(z_j)$ at age 50 conditional on $z = z_j$ as

$$\begin{aligned}
\ell^h_{50}(z_j) &= \sum_{a\in A}\Big[a \sum_{h\in H, z\in Z} [1-\gamma_a(z,h)]\, x_a(z,h)\Big] + 1, \\
x_{a+2}(z',h') &= \sum_{h\in H, z\in Z} p_a(h',z'|h,z)\ \gamma_a(z,h)\ x_a(z,h) \qquad \forall z' \in Z,\ \forall h' \in H,\ \forall a \geq 50, \\
x_{50}(z_j,h) &= \varphi_{50}(h|z_j) \text{ and } x_{50}(z,h) = 0,\ \ \forall z \neq z_j.
\end{aligned}$$

The statistic $\ell^h_{50}(z_j)$ shall be interpreted as the expected remaining life of a given cohort of individuals with characteristic $z = z_j$ at age 50 that face the same age-dependent mortality rates $\gamma_a(z,h)$ conditional on type z and health h, and the same age-dependent joint evolution of type-z and health $p_a(h',z'|h,z)$ as the current old, but may differ on the initial conditional distribution of health $\varphi_{50}(h|z_j)$. Therefore, the socioeconomic gradient of expected longevities computed with use of the health information, $\ell^h_{50}(z_1) - \ell^h_{50}(z_n)$, differs from the one computed in Section 4.1, $\ell_{50}(z_1) - \ell_{50}(z_n)$, by allowing the type-$z$ mortality rates $\gamma_a(z)$ and the law of motion for z, $p_a(z'|z)$, to be different in the future. As discussed above, it does so through the observed differences in the health distribution by type z at age 50 across cohorts.

In columns 2-8 of Table 4, we report the differences in expected longevities between individuals with different socioeconomic characteristics. In all cases, the differentials computed taking into account the information on health are larger than the ones computed without it (see Table 2 for comparison). For instance, the expected longevity differential for males due to education is 6.6 years when using the information on health (whereas it is only 6.3 years when not using it); for wealth it is 4.3 years (3.6); for income it is 1.5 years

it imposes less structure than an ordered logit by allowing the marginal effect of any variable z into health tomorrow to differ from its marginal effect on mortality. This distinction is important. For instance, the effect of education on mortality is null after controlling for health, but it is still an important determinant of the law of motion of health (see Appendices B.1 and B.2). The decompositions in Section 6 are based precisely on this distinction.

(0.8); for labor market status it is 3.8 years (2.0); for marital status it is 2.6 years (2.2); and for smoking it is 2.9 years (2.2). This suggests the existence of significant differences in the distribution of health across cohorts: for the cohorts born 1941-54, the gap in health at age 50 between the least and the most advantaged types was larger than the gap of older cohorts. This implies that the health condition among the least advantaged types of the current old is better than it will be in the future when the 1941-54 cohorts age.

5.3 Summary

We use information on self-rated health to estimate transitions of the socioeconomic characteristics allowing for the joint dependence of socioeconomic characteristics and health and to calculate age specific mortality rates of different health and socioeconomic groups implying a more accurate assessment of the role of those characteristics. We have not found further improvements in average mortality rates in the near future and, in this sense, using expected longevities delivers no gains relative to the use of period life expectancies. If there were to be any improvement in life expectancies in the coming years, it would need to come from reductions of mortality conditional on health. However, the use of health information implies larger estimates of the socioeconomic gradient of expected longevities. This has the added implication that we expect the differences in mortality rates across socioeconomic groups to increase in the coming years. Unfortunately, the stringent data requirements that are needed to use health information prevents us from calculating time trends and expected longevity differentials across finer partitions of the population such as jointly education and socioeconomic characteristics.

6 Decomposing the socioeconomic gradient in expected longevity

We exploit information about health to decompose the expected longevity gradients into three elements: (a) differences in health already present at age 50, (b) changes in health that developed after age 50, and (c) differences in two-year mortality rates unrelated to measured health. To perform this decomposition, we build expected longevities where only one of three elements is allowed to depend on h. That is, instead of the triplet $\{\varphi_{50}(h|z), p_a(z', h'|z, h), \gamma_a(z, h)\}$, we use only one element in turn combined with the other two elements of $\{\varphi_{50}(h), p_a(h'|h), \gamma_a(h)\}$. The results in Table 4 show that, when we look at education, around one-fourth of the life expectancy differential is due to differences in health at age 50 for different education groups, $\varphi_{50}(h|z)$ (see rows (a) for both males and females). That is, college graduates report better self-rated health than individuals without a high school degree. Given the average evolution of health over life, $p_a(h'|h)$, and the average mortality rates by health types, $\gamma_a(h)$, this difference in the initial distribution of health generates by itself a difference in life expectancy of 1.7 years for males and 1.1 for females. The education-specific health transition matrix, $p_a(z', h'|z, h)$, accounts for about three-quarters of the education gap (see rows (b) for both males and females). That is, the fact that self-rated health deteriorates less for highly educated individuals generates by itself a life expectancy gap of 5.0 years for males and 4.7 for females. Finally, the effect of education-specific mortality rates $\gamma_a(z, h)$ is almost null: 0.0 years for males, and 0.3 years for females but with a standard error of 0.5 (see rows (c)). Indeed, in the underlying logit regressions, the effect of education on mortality rates is not statistically different from zero once we control for health; see Appendix B.1 for details. The decomposition for the life expectancy gaps by wealth generates a similar pattern, albeit less dramatic pattern, in which initial health differences account for roughly one-third, the health-protecting nature of wealth accounts for about half, whereas mortality rates account for less than a quarter. In the case of income, the initial health distribution accounts for about half of the advantage of the top quintile, while the rest is shared by the health-protecting nature of income and by mortality specific rates, with the share of the latter being smaller.

The decomposition results are substantially different for smoking and marital status in that mortality rates are smaller for married or nonsmokers even after controlling for self-rated health. The same is true when we look at both types together. In particular, type-specific mortality rates account for between 1/3 and 1/2 of the life expectancy differential for smoking and marital status for both males and females.

Finally, for the labor market status we observe that the initial distribution of health is very important: it accounts for 2.3 years out of 3.8 for males and 0.8 out of 1.4 for females. This is consistent with the evidence that early retirement / inactivity is very much linked to health status.

The difference between education, wealth and income on the one hand, and marital status and smoking on the other point to the limits of socioeconomic privilege. While financial resources may be health protecting, they lower mortality very little conditional on health (nothing in the case of education). We find that, once health has fallen with some terminal condition, financial resources help very little. We do find, however, some advantages in terms of mortality conditional on health related to being married and a non-smoker.

7 Conclusions

We have developed a new methodology to compute differences in the expected longevity of individuals that have different socioeconomic characteristics—education, wealth, non-financial income, labor market attachment, marital status, and smoking status—at age 50. Our measure deals with the two main problems associated to the use of life expectancies: that people's characteristics evolve over time, and that there are time trends in mortality. Our methodology borrows from the literature on duration analysis: we estimate a hazard model for survival with time-varying stochastic endogenous covariates and use it to compute expected durations.[12]. Since the expected life length after age 50 is much longer than our window of observation, we overcome the right-censoring problem by using data for individuals from older cohorts and exploiting information on health h. An ideal alternative way to compute longevity differentials would be by following a cohort of individuals over time until they die. In this manner, the right-censoring problem would be completely eliminated. The problem with this approach is that it requires data that are not available neither in the United States nor in most other countries.

We have uncovered a large amount of heterogeneity in expected longevities, for instance, a man in the top wealth (income) quintile lives 4.3 (1.5) more years than a man in the lowest wealth (income) quintile. However, a naive application of the methodology based on period life expectancies that does not address these two problems yield much larger values: 10.7 (6.1) years for a 50 year old male in the top wealth (income) quintile versus another in the bottom quintile. Our methodology clearly gives much better answers when we are interested in comparing the expected duration of the lives of particular groups of people. Furthermore, the methodology is also applicable to various other countries that have produced data sets with similar information to the one that we have used, like the ELSA (English Longitudinal Study of Ageing) in the United Kingdom or the SHARE (Survey of Health Ageing and Retirement in Europe) in continental Europe. Thus, our approach allows comparisons of the socioeconomic gradient of longevity between different countries, which can nicely complement the scarce international evidence

[12] See Lancaster (1990) for an overview of duration analysis

on the socioeconomic gradient of life expectancies, see Majer *et al.* (2010).

This large amount of heterogeneity in expected life duration matters for a number of reasons. First, the socioeconomic gradient in expected longevity probably dwarfs the welfare implications of the income differences accruing to different socioeconomic groups. Second, the redistributive power of public policies that are paid out as life annuities—such as retirement pensions, public medical assistance, or long-term care—may be partly eroded by the longer life expectancies of richer individuals. For instance, Fuster *et al.* (2003) show that the life expectancy differences between education groups makes the social security system more beneficial to the highly educated, despite the strong redistributive component introduced into the system. Finally, financial products such as life annuities, life insurance, or medical insurance are intimately related to the expected length of life. The measurement of expected longevities of different population subgroups may help assess whether the pricing of these products is actuarially fair, which is itself important to understand the take up of these products.

Decomposing the differences in longevity into a fraction that is due to differences in self perceived health at age 50, changes in health after age 50 and differences in mortality among groups with different socioeconomic characteristics shows that by far and large the most important component is the advantage that various socioeconomic groups have in preserving health. A salient finding here is that, while education and wealth seem to have little predictive power for two-year survival rates once self assessed health is known, marital status does help predict survival. This raises important questions for further research in trying to understand the survival advantage of married people.

We also document an increasing time trend of the differentials among socioeconomic groups during the sample period 1992-2008. As it is well known, income and wealth inequality have also risen during this same period. We do not know whether these two phenomena are connected, but it is certainly worth exploring. At the same time, our results show that the socioeconomic gradients of expected longevities are larger when we use information on self assessed health in order to predict the future mortality rates and the future transition matrices between types of the current young. This implies that the socioeconomic differences in mortality are likely to widen in the coming years.

Finally, it is important to highlight that our methodology has also its drawbacks. First, the Markov assumption, both in the transitions and the mortality rates, is certainly restrictive and rules out effects from the distant past. However, adding more lags to the estimates has also problems. Every added lag would require longer spells in the panel dimension and hence it would diminish the effective sample size. In addition, since states are quite persistent, the identification of the effects of every lag would be difficult, and the precision of the estimates would suffer. Second, any measurement error in socioeconomic characteristics would generate artificially high mobility between states. This would make the difference between the socioeconomic gradients in life expectancy and expected longevity artificially large. Hence, our method should be applied to data sets were socio-economic status is measured with confidence. And third, the HRS and related data sets do not provide large samples, which generate relatively large standard errors, in particular when looking at characteristics with many states.

Appendix

A Data

We use version M of the RAND files of the HRS, which covers 10 waves from 1992 to 2010.

A.1 Variable definitions

Education. Variable RAEDUC provides five educational categories: no high school degree, high school dropout with GED tests, high school graduate, high school graduate with some college, and college graduate. We pool the second, third, and fourth categories together for a wider high school graduate category.

Wealth. The HRS provides several measures of assets and liabilities. We define wealth as total household net worth per adult, excluding second residences and mortgages on second residences. The reason being that these two variables are available neither for the third wave of the whole sample nor for the second wave of the AHEAD subsample. Hence, total wealth is the sum of HwASTC, HwACHCK, HwACD, HwABOND, HwAOTH, HwAHOUS, HwARLES, HwATRANS, HwABSNS, and HwAIRA, minus HwDEBT, Hw-MORT, and HwHMLN. Then, we divide the resulting figure by 2 if the individual is married. We deflate our resulting variable by the CPI. Finally, in order to have a discrete version of the wealth variable, we classify every individual-year observation by the quintile of that individual-year observation in the wealth distribution over all individual-year observations, including both white males and white females. Hence, the top quintile represents individuals with household wealth over \$207,450 the second quintile has a lower cutoff point of \$95,497, the third quintile has a lower cutoff point of \$44,677, and the fourth quintile has a lower cutoff point of \$11,597. All figures are in 1992 dollars.

Income. We use information on non-financial income, which is measured at the individual level. This is the sum of RwIEARN (labor earnings), RwIUNWC (unemployment benefits), RwIPENA (employer pensions and annuities) RwISRET (retirement income from social security) RwISSDI (disability income from social security), and RwIGXFR (other government transfers). We deflate our resulting variable by the CPI and we compute the quintile where the individual-year observation belongs in the income distribution over all individual-year observations, including both white males and white females. Hence, the top quintile represents individuals with non-financial income over $25,804 the second quintile has a lower cutoff point of $14,096, the third quintile has a lower cutoff point of $8,418, and the fourth quintile has a lower cutoff point of $4,554. All figures are in 1992 dollars.

Labor market status. Variable RwLBRF provides seven categories for the relationship of the respondent with the labor market. We reduce it to three: attached, semi-attached and inactive. In the first category we include individuals that are either working full-time or unemployed and looking for a job; in the second category we include people working part-time or semiretired; in the third we include individuals who are retired, disabled, or out of the labor force.

Marital status. Variable RwMSTAT and classify as married those who answer to be either married or partnered, and classify as nonmarried the rest: separated, divorced, widowed, and never married.

Smoking. Variable RwSMOKEN that reports whether the respondent is currently a smoker.

Health. Variable RwSHLT, which reports five categories (excellent, very good, good, fair, and poor) for the respondent's self-reported general health status.

Alive. For every individual-year observation we need to determine whether he/she survives into the next wave. Every wave contains the variable RwIWSTAT that gives the response and mortality status of the respondent. Code 1 indicates that the respondent actually responded to the interview, so he/she is alive. Code 4 indicates that the respondent dropped from the sample, but a follow-up on him could be done and it was verified that he/she was alive. These two cases are the ones we count as alive. Code 5 indicates that the individual did not make it alive to the current wave. Finally, code 7 indicates that the individual withdrew from the sample (due to either the sample design or sample attrition) and his/her survival is not known. We classify these cases as 'missing'.

A.2 Sample selection

We start with all individuals in the age range 50 to 94. For every individual-year observation we record the relevant information for the next wave and drop individual-year observations in 2010 —because information for the next wave is not available— and individual-year observations with zero sampling weight. This yields 12,219 males and 15,081 females. We drop individuals with missing information for race (18 males and 17 females), nonwhite individuals (2,634 males and 3,351 females), and individuals with missing information for education (2 males and 1 female), which leaves 9,836 males and 11,713 females. Every individual is observed in several waves, and every individual-year observation is useful to estimate survival probabilities and transitions of our covariates. Overall, we have 59,167 and 74,372 individual-year observations for males and females. We next drop individual-year observations for which we do not know survival status into the next wave (596 and 748 individual-year observations). This happens for some observations of people who could not be followed upon withdrawing from the sample (code 7 in RwIWSTAT). Next, we drop individual-year observations with missing information on health (5,209 for males and 6,171 for females, which correspond to 286 and 331 different individuals). What is left is our working sample, with 9,542 males that provide 53,362 individual-year observations and 11,236 females that provide 67,453 individual-year observations. Out of these individual-year observations, we have 3,431 deaths for males and 3,235 for females, with average death rates of 6.4% and 4.8%.

B Estimation of the underlying logistic regressions

B.1 Survival probabilities

In Sections 3 and 4, we approximate parametrically the survival probabilities γ_a as a function of age a only, and $\gamma_a(z)$ as a function of age and some type $z \in Z$. We run logistic regressions of survival as follows:

$$\text{Prob}\left(alive_{t+2} = 1 | a_t, z_t\right) = \frac{e^{f(a_t,z_t)}}{1 + e^{f(a_t,z_t)}}$$

and

$$f\left(a_t, z_t\right) = \alpha_0 + \alpha_1 a_t + \sum_{i=2}^{I} \alpha_{2i} D_{z_t=z_i} + \sum_{i=2}^{I} \alpha_{3i} \left(D_{z_t=z_i} \times a_t\right),$$

where $D_{z_t=z_i}$ is a dummy variable that takes value one if $z_t = z_i$ and zero otherwise, and $alive_{t+2}$ is a dummy variable that takes value one if the individual is alive in the next wave and zero otherwise. In Table 5 we show the results of these regressions for white males. The results of other regressions are available upon request. The categories into the set Z are always sorted from the most to the least advantaged type. We also tried specifications that add quadratic or cubic terms on age. In these specifications the quadratic term is significant and improves the fit slightly. However, it does not change the computed life expectancies. Later on, when we keep adding variables to the regression and interact them with age, it helps to have a parsimonious specification. In Figure 2 we plot the survival rates for white males against age. In panel (a) we plot the survival rates with the age term obtained through age dummies, and then also a linear, a quadratic, and a cubic polynomial. We see how the quadratic polynomial improves the linear one in that it better captures the fall in survival in the very last years. In panel (b) we plot the survival for college-educated males and individuals without a high school degree with the age term captured either through age dummies or through a linear term, in both cases interacted with education. We see that the linear term is enough to capture the shape of the age profile in both education cases. In panels (c) and (d) we do the same for health, and again the linear term does very well in capturing the different shapes of each health group.

[Table 5 about here.]

[Figure 2 about here.]

In Section 4 we also look at the expected longevity differentials in different years, and hence we need to compute time-dependent age-specific survival probabilities $\gamma_{a,t}$ and $\gamma_{a,t}(z)$. To do so, we include a variable t for calendar year, as well as its interaction with the type variable z. We do not, however, interact it with age. See footnote 6 for a discussion.

$$\text{Prob}\left(alive_{t+2} = 1 | t, a_t, z_t\right) = \frac{e^{f(t,a_t,z_t)}}{1 + e^{f(t,a_t,z_t)}}$$

and

$$f\left(t, a_t, z_t\right) = \alpha_0 + \alpha_1 a_t + \sum_{i=2}^{I} \alpha_{2i} D_{z_t=z_i} + \sum_{i=2}^{I} \alpha_{3i}\left(D_{z_t=z_i} \times a_t\right) + \alpha_4 t + \sum_{i=2}^{I} \alpha_{5i}\left(D_{z_t=z_i} \times t\right).$$

We do not report the results of these and the following regressions, but they are available upon request.

In Section 4 we also look at the expected longevities by education group. For this exercise we run the same survival regressions but for the given education subpopulation only.

In Section 5 we compute survival probabilities $\gamma_a\left(h\right)$ and $\gamma_a\left(h, z\right)$, where $h \in H$ is self-rated health. We use the same logistic regression upgraded to include health:

$$\text{Prob}\left(alive_{t+2} = 1 | a_t, h_t\right) = \alpha_0 + \alpha_1 a_t + \sum_{j=2}^{J} \alpha_{2j}\left(D_{h_t=h_j}\right) + \sum_{j=2}^{J} \alpha_{3j}\left(D_{h_t=h_j} \times a_t\right)$$

and

$$\begin{aligned} \text{Prob}\left(alive_{t+2} = 1 | a_t, z_t, h_t\right) = \alpha_0 + \alpha_1 a_t \quad & + \quad \sum_{j=2}^{J} \alpha_{2j}\left(D_{h_t=h_j}\right) + \sum_{j=2}^{J} \alpha_{3j}\left(D_{h_t=h_j} \times a_t\right) \\ & + \quad \sum_{i=2}^{I} \alpha_{4i}\left(D_{z_t=z_i}\right) + \sum_{i=2}^{I} \alpha_{5i}\left(D_{z_t=z_i} \times a_t\right) \end{aligned}$$

An important finding in this paper is that, after controlling for self-rated health, differences in educational attainment have very little predictive power for two-year-ahead mortality rates. To see this, we first look at the predictive power of each variable alone. In Figure 3 we plot the predicted survival rates by health—panel (a)—and education—panel (b)—when the logistic regressions include only health or education variables. The clear result here is that differences in self-rated health imply much larger differences in survival than do differences in education.

[Figure 3 about here.]

When we put both types of variables together in the same regression, we can use differences in the likelihood to test how much information education adds to health and how much information health adds to education. In Table 6 we report some results from the three logistic regressions: only health variables, only education variables, and both together. The first row corresponds to survival depending only on education, the second row to survival depending only on health, and the third row to survival depending on both. As suggested by Figure 3, the odds ratios are much larger when comparing education categories than when comparing health categories. Interestingly, when adding the two types of variables the odds ratios of education become statistically not different from one. Instead, the odds ratios for health become slightly larger when adding education to the regression. The results of the likelihood ratio test are clear. Compared with the model regression with both education and health, the constraint that all the coefficients on the health variables are zero is rejected strongly. Instead, the constraints that the education variables are zero is rejected at the 10% confidence level but not at 1%.

[Table 6 about here.]

A visual inspection of these results comes from Figure 4. In panel (a) we reproduce the survival rates by education group as in Figure 3, panel (b). In panels (b), (c), and (d) we plot the predicted survival rates by education group within a health category. It is easy to see that within the health category, the role of education is minimal.

[Figure 4 about here.]

B.2 Transition functions

In Section 4, we compute transition matrices $p_a\left(z'|z\right)$ by multivariate logistic regressions as follows:

$$\begin{aligned}\text{Prob}\left(z_{t+2}=z_1|a_t,z_t\right) &= \frac{1}{1+\sum_{j=2}^{I}e^{f_j(a_t,z_t)}}\\ \text{Prob}\left(z_{t+2}=z_k|a_t,z_t\right) &= \frac{e^{f_k(a_t,z_t)}}{1+\sum_{j=2}^{I}e^{f_j(a_t,z_t)}} \qquad \forall 1<k\leq I\end{aligned}$$

with

$$f_k\left(a_t,z_t\right)=\alpha_{k0}+\alpha_{k1}a_t+\sum_{i=2}^{I}\alpha_{k2i}D_{z_t=z_i}+\sum_{i=2}^{I}\alpha_{k3i}\left(D_{z_t=z_i}\times a_t\right).$$

In Section 4 when we need to compute time-dependent transition matrices $p_{a,t}(z'|z)$, we add a variable for calendar year and interact it with the dummies for type:

$$\begin{aligned}
\text{Prob}\left(z_{t+2}=z_1|t,a_t,z_t\right) &= \frac{1}{1+\sum_{j=2}^{I} e^{f_j(t,a_t,z_t)}} \\
\text{Prob}\left(z_{t+2}=z_k|t,a_t,z_t\right) &= \frac{e^{f_k(a_t,z_t)}}{1+\sum_{j=2}^{I} e^{f_j(t,a_t,z_t)}} \qquad \forall 1<k\leq I
\end{aligned}$$

with

$$f_k\left(t,a_t,z_t\right)=\alpha_{k0}+\alpha_{k1}a_t+\sum_{i=2}^{I}\alpha_{k2i}D_{z_t=z_i}+\sum_{i=2}^{I}\alpha_{k3i}\left(D_{z_t=z_i}\times a_t\right)+\alpha_{k4}t+\sum_{i=2}^{I}\alpha_{k5i}\left(D_{z_t=z_i}\times t\right).$$

Finally, in Section 5 when we need to compute transition matrices $p_a(h'|h)$ and $p_a(z',h'|z,h)$, we follow a similar approach. In the first case, we replace the $z\in Z$ by $h\in H$. In the second one we create new dummy variables by combining $Z\times H$. This implies estimating very large models: the case for assets requires 25 outcome variables (5 asset categories times 5 health types). An alternative for the transition matrices for the self-rated health would be to use an ordered logit. This approach is attractive because by imposing the structure of the ordered logit, we need to estimate much fewer parameters, and hence we could potentially add more variables together. However, the restrictions imposed by the ordered logit are statistically rejected, so we stay with the multivariate logit.

B.3 Standard errors

We compute the standard errors of the life expectancies and the expected longevities by simulation. We follow Diermeier *et al.* (2003) and draw 25,000 vectors of parameters from the corresponding estimated asymptotic distribution, and compute the desired life expectancies or expected longevities with each vector or parameters. This gives us a sample of 25,000 observations for each statistic of interest, and we report the mean and the standard deviation of this distribution. The vectors of parameters estimated in the logistic and multivariate logistic regressions are asymptotically normally distributed, with the mean given by the point estimates, and the variance-covariance matrix of the parameters also obtained from the estimation process. The initial distributions $\varphi_{50}(h)$ and $\varphi_{50}(h|z_j)$ that we use in Sections 5 and 6 follow multinomial distributions.

References

Attanasio, O. and Hoynes, H. (2000). Differential mortality and wealth accumulation. *Journal of Human Resources*, **35** (1), 1–29.

Brown, D. C., Hayward, M. D., Montez, J. K., Hummer, R. A., Chiu, C.-T. and Hidajat, M. M. (2012). The significance of education for mortality compression in the united states. *Demography*, **49** (3), 819–840.

Brown, J. (2002). Differential mortality and the value of individual account retirement annuities. In M. Feldstein and J. B. Liebman (eds.), *The Distributional Aspects of Social Security and Social Security Reform*, *10*, University of Chicago Press.

Crimmins, E. M., Hayward, M. D. and Seeman, T. (2004). Race/ethnicity, socioeconomic status and health. In N. B. Anderson, R. A. Bulatao and B. Cohen (eds.), *Critical Perspectives on Racial and Ethnic Differences in Health in Later Life*, Washington, DC: National Academy Press, pp. 310–352.

Currie, J. and Madrian, B. C. (1999). Health, health insurance and the labor market. In O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, vol. 3, *50*, Elsevier Science Publishers, pp. 3309–3416.

Deaton, A. and Paxson, C. (1994). Mortality, education, income and inequality among american cohorts. In D. A. Wise (ed.), *Themes in the Economics of Aging*, *8*, University of Chicago Press, pp. 129–170.

Díaz-Giménez, J., Glover, A. and Ríos-Rull, J.-V. (2011). Facts on the distributions of earnings, income, and wealth in the united states: 2007 update. *Federal Reserve Bank of Minneapolis Quarterly Review*, **34** (1), 2–31.

—, Quadrini, V. and Ríos-Rull, J.-V. (1997). Dimensions of inequality: Facts on the U.S. distribution of earnings, income and wealth. *Federal Reserve Bank of Minneapolis Quarterly Review*, **21** (2), 3–21.

Diermeier, D., Eraslan, H. and Merlo, A. (2003). A structural model of government formation. *Econometrica*, **71** (1), 27–70.

Elo, I. T. and Preston, S. H. (1996). Educational differentials in mortality: United states, 1979-85. *Social Science & Medicine*, **42** (1), 47–57.

Fuster, L., İmrohoroğlu, A. and İmrohoroğlu, S. (2003). A welfare analysis of social security in a dynastic framework. *International Economic Review*, **44** (4), 1247–1274.

Heathcote, J., Perri, F. and Violante, G. (2010). Unequal we stand: An empirical analysis of economic inequality in the united states, 1967- 2006. *Review of Economic Dynamics*, **1** (13), 15–51.

Hu, Y. and Goldman, N. (1990). Mortality differentials by marital status: An international comparison. *Demography*, **27** (2), 233–250.

Idler, E. and Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior*, **38** (1), 21–37.

— and — (1999). Community studies reporting association between self-rated health and mortality. *Research On Aging*, **21** (3), 392–401.

Kitagawa, E. M. and Hauser, P. M. (1973). *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge: Harvard University Press.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge; New York and Melbourne: Cambridge University Press.

Lee, R. and Carter, L. (1992). Modeling and forecasting u.s. mortality. *Journal of the American Statistical Association*, **87** (419), 659–671.

Lin, C., Rogot, E., Johnson, N., Sorlie, P. and Arias, E. (2003). A further study of life expectancy by socioeconomic factors in the national longitudinal mortality study. *Ethnicity and Disease*, **13**, 240–247.

Majer, I., Nusselder, W., Mackenbach, J. and Kunst, A. (2010). Socioeconomic inequalities in life and health expectancies around official retirement age in 10 western-european countries. *Journal of Epidemiology and Community Health*.

Marmot, M. G., Shipley, M. J. and Rose, G. (1984). Inequalities in death–specific explanations of a general pattern? *The Lancet*, **323** (8384), 1003–1006.

—, Smith, G. D., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E. and Feeney, A. (1991). Health inequalities among british civil servants: the whitehall ii study. *The Lancet*, **337** (8754), 1387–1393.

Meara, E., Richards, S. and Cutler, D. (2008). The gap gets bigger: Changes in mortality and life expectancy, by education, 1981? 2000. *Health Affairs*, **27** (2), 350–360.

Montez, J. K., Hummer, R. A., Hayward, M. D., Woo, H. and Rogers, R. G. (2011). Trends in the educational gradient of u.s. adult mortality from 1986 through 2006 by race, gender, and age group. *Research on Aging*, **2** (33), 145–171.

Preston, S. H. and Elo, I. T. (1995). Are educational differentials in adult mortality increasing in the united states? *Journal of Aging and Health*, **7** (4), 476–496.

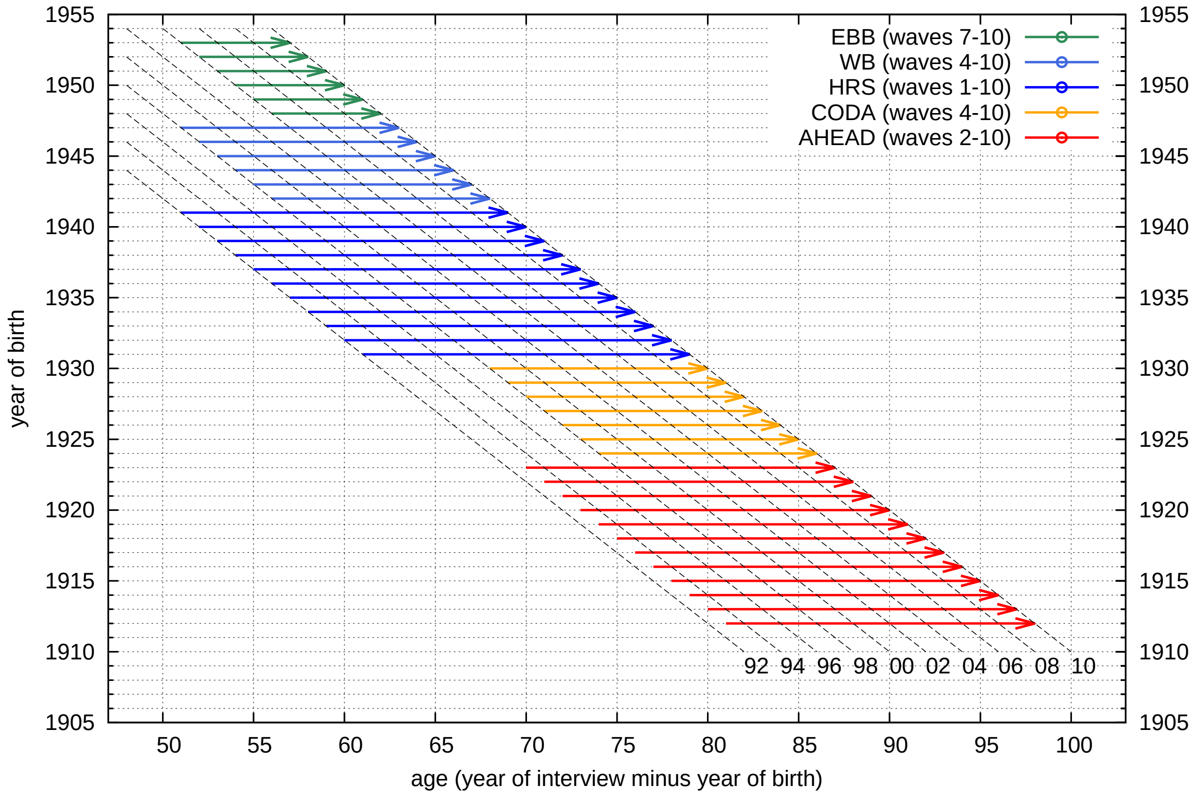
Rogers, R. G., Everett, B. G., Saint-Onge, J. M. and Krueger, P. M. (2010). Social, behavioral, and biological factors, and sex differences in mortality. *Demography*, **47** (3), 555–578.

Singh, G. K. and Siahpush, M. (2006). Widening socioeconomic inequalities in us life expectancy, 1980-2000. *International Journal of Epidemiology*, **35**, 969–979.

Smith, J. P. (1999). Healthy bodies and thick wallets: The dual relationship between health and economic status. *Journal of Economic Perspectives*, **13** (2), 145–166.

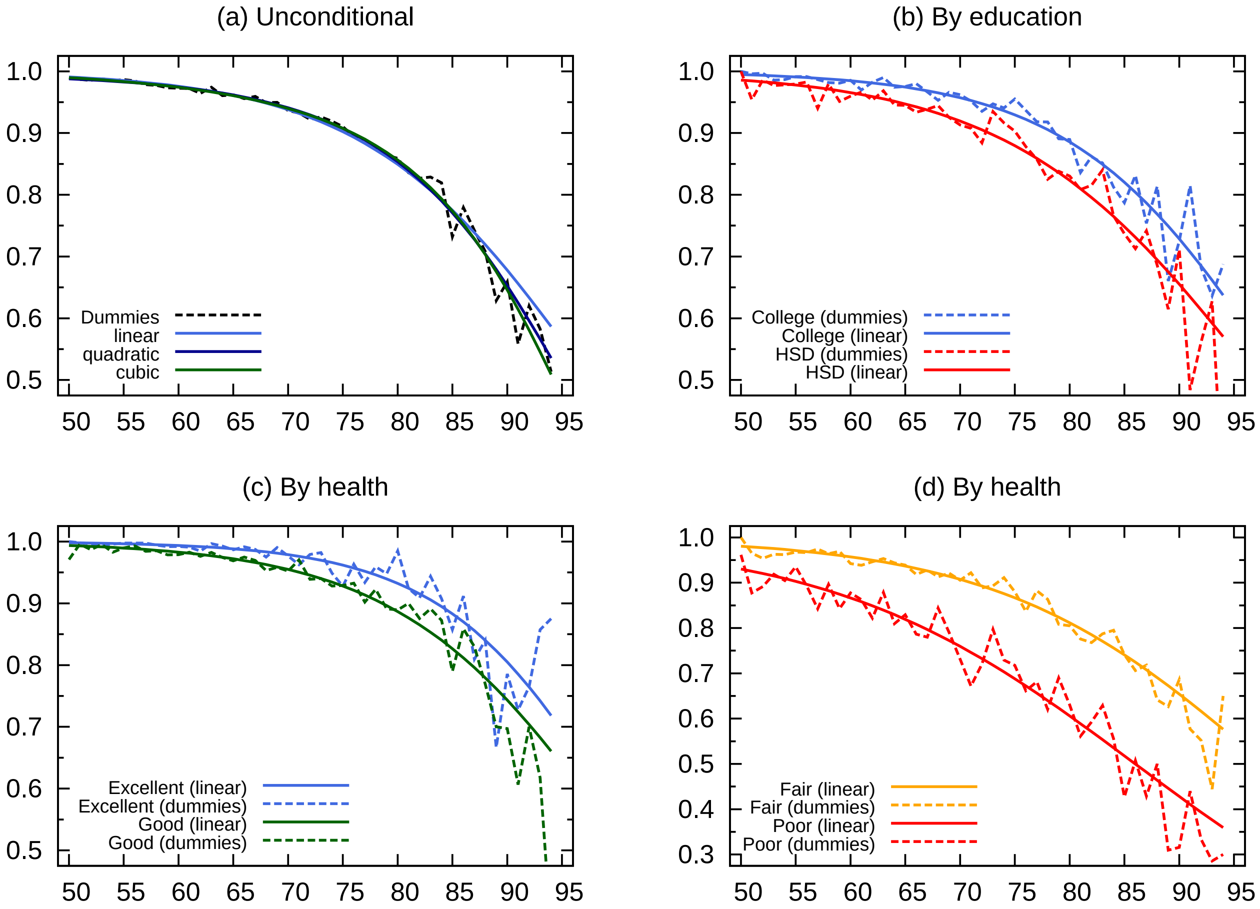
Yogo, M. (2009). Portfolio choice in retirement: Health risk and the demand for annuities, housing, and risky assets, nBER Working Paper 15307.

Figure 1: Age structure of the HRS eligible individuals



Notes: The arrows represent the maximum age range in which eligible individuals of a given year of birth are interviewed. The colors denote different HRS samples. Age on the X-axis is year of interview minus year of birth; actual age may be one year younger.

Figure 2: Survival rates: parametric vs non-parametric age


(a) Unconditional

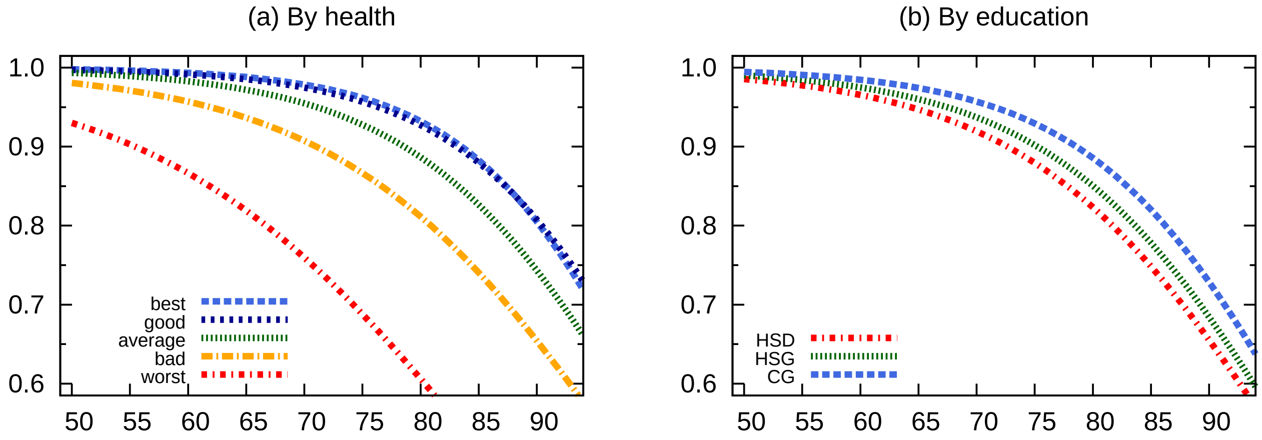
(b) By education

(c) By health

(d) By health

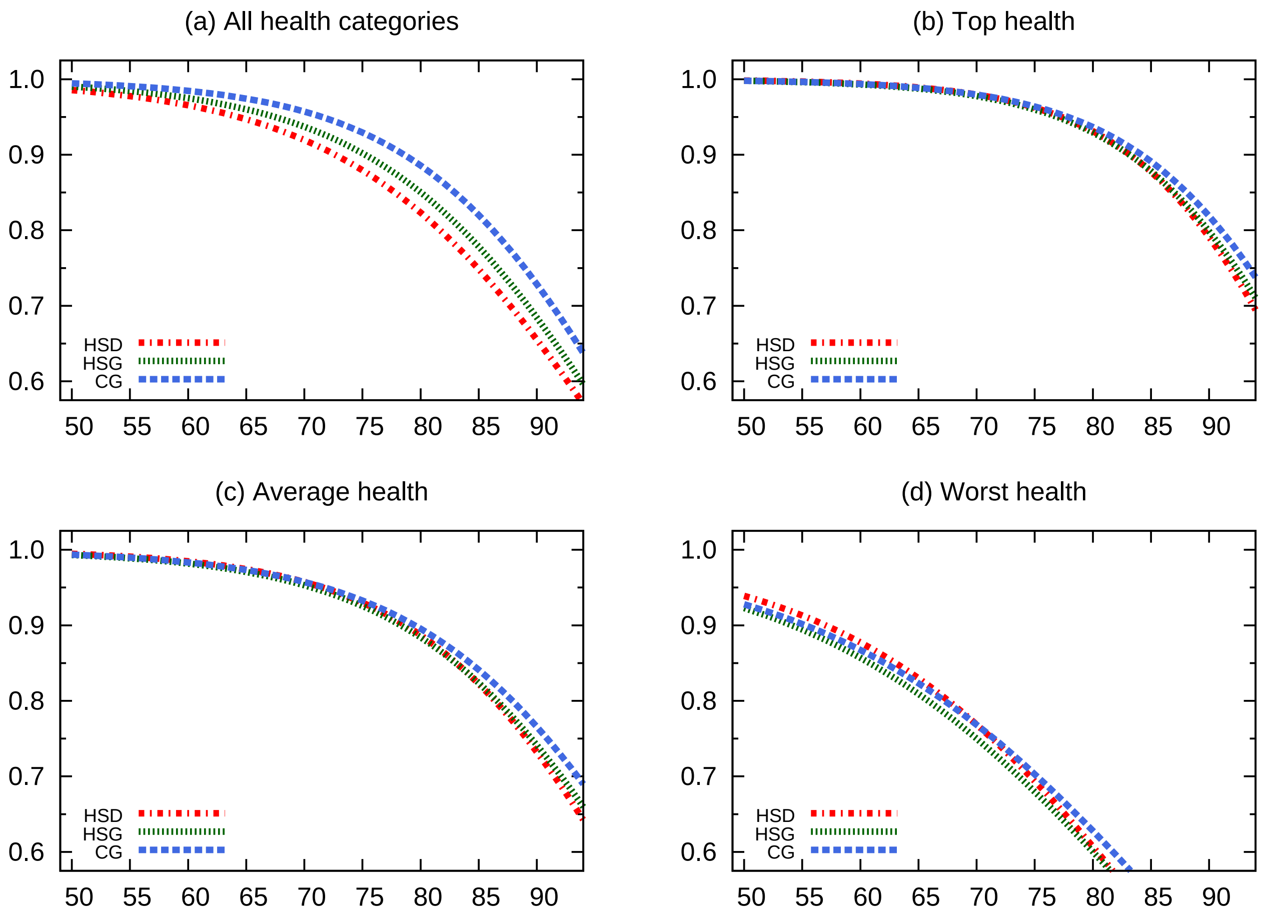
Notes: Predicted yearly survival rates at given age, sample of white males. Since estimates correspond to two-year survivals, we report the squared root of the predictions from our logit regressions. NHSD refers to no high school degree.

Figure 3: Survival rates: by health and by education



Notes: Predicted yearly survival rates at given age, sample of white males. Since estimates correspond to two-year survivals, we report the squared root of the predictions from our logit regressions. NHSD refers to no high school degree, HSG to high school graduates, CG to college graduates.

Figure 4: Survival rates: by health and education jointly



Notes: Predicted yearly survival rates at given age, sample of white males. Since estimates correspond to two-year survivals, we report the squared root of the predictions from our logit regressions. NHSD refers to no high school degree, HSG to high school graduates, CG to college graduates.

Table 1: Period life expectancies at age 50

		Life expectancy gradient						
	All	edu	wea	inc	lms	mar	smok	m-s
White males	78.8 (0.2)	6.3 (0.7)	10.7 (0.8)	6.1 (0.9)	9.4 (1.0)	4.6 (0.6)	7.2 (0.5)	10.2 (0.9)
White females	82.9 (0.2)	5.8 (0.7)	8.5 (0.6)	3.6 (0.7)	5.0 (1.8)	2.4 (0.4)	6.2 (0.5)	7.5 (0.7)

Notes: The first column reports the period life expectancy at age 50 for white males and white females. The remaining columns report the difference in life expectancy between the most and the least advantaged types for education (edu), wealth (wea), non-financial income (inc), labor market status (lms), marital status (mar), smoking (smok), and the combination of smoking and marital status (m-s). See Appendix A for the exact variable definitions. Standard errors in parenthesis.

Table 2: Expected longevity gradients at age 50

	wea	inc	lms	mar	smok	m-s
White males						
(1) All	3.6 (0.4)	0.8 (0.2)	2.0 (0.2)	2.2 (0.3)	2.2 (0.2)	4.6 (0.6)
(2) College graduates	2.8 (0.6)	0.7 (0.2)	1.5 (0.2)	1.6 (0.3)	1.6 (0.2)	3.5 (0.5)
(3) No high school degree	3.7 (0.5)	0.8 (0.2)	2.3 (0.3)	2.9 (0.4)	2.7 (0.3)	5.6 (0.7)
(4) Education and z	8.4 (0.9)	6.5 (0.7)	7.9 (0.8)	8.9 (0.9)	8.0 (0.8)	10.8 (1.1)
White females						
(1) All	2.9 (0.2)	0.8 (0.1)	0.8 (0.1)	1.1 (0.2)	1.8 (0.3)	2.7 (0.3)
(2) College graduates	1.5 (0.3)	0.4 (0.1)	0.4 (0.1)	0.7 (0.1)	0.7 (0.2)	1.3 (0.2)
(3) No high school degree	2.9 (0.3)	0.5 (0.1)	1.0 (0.2)	1.3 (0.3)	2.4 (0.3)	3.5 (0.5)
(4) Education and z	7.2 (0.7)	6.0 (0.7)	6.3 (0.7)	6.9 (0.7)	7.5 (0.7)	8.3 (0.8)

Notes: Average expected longevity differences according to different measures of socioeconomic status. See column labels in footnote of Table 1. Rows (1) refer to all individuals, rows (2) look at the subgroup of college graduates, and rows (3) look at the subgroup of individuals without a high school degree. Rows (4) report the difference between individuals with a college degree and the most advantaged type z and individuals without a high school degree and of the least advantaged type z. Standard errors in parenthesis.

Table 3: Time trends in expected longevities at age 50

		Expected longevity gradients						
	All	edu	wea	inc	lms	mar	smok	m-s
White males								
1992	77.6 (0.3)	5.1 (1.4)	2.4 (0.4)	0.9 (0.2)	1.6 (0.2)	1.5 (0.4)	1.6 (0.3)	3.6 (0.8)
2008	79.6 (0.5)	6.8 (1.1)	4.5 (0.7)	0.8 (0.2)	2.3 (0.3)	2.7 (0.5)	2.7 (0.4)	5.4 (0.8)
Δ	+2.1 (0.7)	+1.7 (2.1)	+2.1 (0.8)	–0.1 (0.2)	+0.7 (0.3)	+1.1 (0.6)	+1.1 (0.5)	+1.8 (1.2)
1992 (*NVSS*)	77.1							
2008 (*NVSS*)	79.3							
Δ (*NVSS*)	+2.2							
White females								
1992	83.0 (0.4)	4.6 (1.3)	2.5 (0.3)	0.8 (0.1)	0.4 (0.2)	0.6 (0.4)	1.1 (0.3)	1.4 (0.5)
2008	82.8 (0.3)	7.0 (1.0)	3.1 (0.3)	0.7 (0.1)	1.1 (0.2)	1.5 (0.3)	2.4 (0.5)	3.9 (0.6)
Δ	–0.3 (0.6)	+2.4 (2.0)	+0.6 (0.4)	–0.2 (0.2)	+0.7 (0.2)	+0.9 (0.5)	+1.3 (0.6)	+2.6 (0.9)
1992 (*NVSS*)	81.9							
2008 (*NVSS*)	82.9							
Δ (*NVSS*)	+1.0							

Notes: The first column reports the period life expectancy at age 50 for different years. The remaining columns report the average expected longevity differences according to different measures of socioeconomic status for different years. See column labels in footnote of Table 1. Δ refers to the difference between 1992 and 2008. The first three rows in each panel refer to our own calculations with the HRS, while the next three lines refer to the data reported by the NVSS. Standard errors in parenthesis.

Table 4: Expected longevities at age 50 with health status

		Expected longevity gradients						
	All	edu	wea	inc	lms	mar	smok	m-s
White males								
All type-specific	78.8 (0.2)	6.6 (0.8)	4.3 (0.6)	1.5 (0.4)	3.8 (0.4)	2.6 (0.6)	2.9 (2.4)	5.8 (2.7)
(a) only initial health		1.7 (0.2)	1.4 (0.2)	0.8 (0.2)	2.3 (0.3)	0.5 (0.2)	0.6 (0.1)	1.1 (0.3)
(b) only transition		5.0 (0.4)	2.0 (0.4)	0.4 (0.2)	0.5 (0.1)	0.8 (0.5)	1.0 (1.9)	1.8 (2.6)
(c) only mortality		0.0 (0.6)	1.0 (0.3)	0.3 (0.2)	0.7 (0.2)	1.2 (0.3)	1.2 (0.2)	2.7 (0.4)
White females								
All type-specific	83.0 (0.2)	5.8 (0.5)	3.7 (0.4)	1.3 (0.2)	1.4 (0.2)	1.4 (0.5)	2.2 (0.3)	3.2 (1.4)
(a) only initial health		1.1 (0.0)	1.1 (0.1)	0.7 (0.1)	0.8 (0.1)	0.3 (0.1)	0.2 (0.1)	0.4 (0.1)
(b) only transition		4.7 (0.3)	1.9 (0.2)	0.5 (0.1)	0.3 (0.1)	0.8 (0.4)	1.0 (0.1)	1.6 (1.2)
(c) only mortality		0.3 (0.5)	0.7 (0.2)	0.2 (0.1)	0.3 (0.1)	0.3 (0.1)	0.9 (0.2)	1.1 (0.2)

Note: The first column reports the expected longevity at age 50 for white males and white females, making use of the background information on health. The remaining columns report the difference in expected longevities between the most and the least advantaged types, making use of the background information on health. See column labels in footnote of Table 1. Rows (a), (b), and (c) measure the contribution of differences in health at age 50, differences in the evolution of health after age 50, and differences in mortality rates across socioeconomic groups for the overall gradient. Standard errors in parenthesis.

Table 5: Logistic regressions for survival

	edu	wea	lms	mar	smok	m-s
constant	10.53***	10.89***	9.015***	9.755***	10.48***	10.46***
	(31.28)	(30.91)	(17.62)	(59.56)	(65.12)	(55.31)
$D_{z_t=z_2}$	-1.075***	-0.246	0.252	-1.565***	-1.954***	-1.413***
	(-2.78)	(-0.51)	(0.31)	(-5.22)	(-5.50)	(-3.17)
$D_{z_t=z_3}$	-1.810***	-1.134**	-1.269**			-1.037***
	(-4.33)	(-2.43)	(-2.35)			(-2.35)
$D_{z_t=z_4}$		-1.475***				-3.368***
		(-3.24)				(-6.18)
$D_{z_t=z_5}$		-2.875***				
		(-6.47)				
age	-0.106***	-0.109***	-0.0791***	-0.0991***	-0.109***	-0.107***
	(-23.72)	(-23.94)	(-9.68)	(-45.12)	(-51.96)	(-43.00)
age$\times D_{z_t=z_2}$	0.010*	-0.00031	-0.00460	0.0148***	0.0169***	0.00991
	(1.86)	(-0.05)	(-0.38)	(3.79)	(3.32)	(1.54)
age$\times D_{z_t=z_3}$	0.016***	0.00969	0.00351			0.00881*
	(2.96)	(1.58)	(0.41)			(1.84)
age$\times D_{z_t=z_4}$		0.0107*				0.0315***
		(1.79)				(4.02)
age$\times D_{z_t=z_5}$		0.0232***				
		(3.99)				
N	56322	56299	55590	56299	56037	56014

Notes: Sample of white males. See column labels in footnote of Table 1. The omitted dummy variable corresponds to the most advantaged category ($z_t = z_1$), and the rest of the dummies are ordered towards the least advantaged categories. t-statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Logistic regressions for survival: health and education

	Odds ratios				**LR test**	
	CG **vs** $NHSD$		H_2 **vs** H_4		χ^2	**p-value**
	age 65	**age 75**	**age 65**	**age 75**		
Only education	4.46	4.30			1897.89	0.000
Only health			171.19	170.90	9.32	0.054
Both together	1.08	1.16	202.29	202.02		

Notes: Sample of white males. Each row corresponds to a logistic regression of survival against a linear term in age, the corresponding dummies (education dummies in the first row, health dummies in the second row, both sets of dummies in the third row), and the interaction of the linear term in age and the dummies. $NHSD$ and CG refer to no high school degree and college graduates respectively. H_2 and H_4 correspond to the second best and second worst health categories respectively. The LR (likelihood ratio) tests in the first and second row test the corresponsing models against the model in the third row. Therefore, the null hypothesis in the first row is that all the health variables are zero and that only education matters (which is clearly rejected); and the null hypothesis in the second row is that all the education variables are zero and that only health matters (which is marginally rejected).