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

September 5, 2012

Abstract

We develop a new methodology to compute differences in the expected longevity of individuals who are in different socioeconomic groups at age 50. We deal with the two main problems associated with the standard use of life expectancy: that people's socioeconomic characteristics evolve over the life cycle and that there is a time trend that reduces mortality over time. Using HRS data we uncover an enormous amount of heterogeneity in expected longevities between individuals in different socioeconomic groups. Additionally, our analysis allows us to provide an answer to the old question of how health protecting are education, wealth and marital status. To do so, 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, the latter is the least important for most socioeconomic characteristics. In particular, education and wealth are health protecting but have very little impact on two-year mortality rates conditional on health. Finally, we document an increasing time trend of the socioeconomic gradient of longevity in the period 1992-2008, and a likely increase in the socioeconomic gradient of mortality rates in the near future. Last but not least, we show that the longevity differences that we find have welfare implications that dwarf the differences in consumption accruing to people in different socioeconomic groups.

JEL classification: I14; I24; J12; J14

Keywords: Inequality in health; Heterogeneity in mortality rates; Life expectancies

[*] We are grateful for comments from seminars at Penn and Carlos III, as well as those from a LAEF conference on health and macroeconomics. 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 falling with years of 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 affect the total duration of people's lives, which in the end is what we really want to know. Consider, for instance, the case of a policymaker who needs to predict the future path of public spending in retirement pensions. Or consider the case of a financial actor who wants to price a life annuity contract. Yet, no clear methodology to predict expected life duration of different population subgroups is available. One possibility is to use differences in life expectancies, which mechanically aggregate socioeconomic differences in the mortality rates from a given calendar year.[2] The concept of *life expectancy*, however, may be very different from the notion of *expected longevity*, which is the concept that we are interested in. There are two reasons for this difference. The first of these reasons is that people's socioeconomic characteristics evolve over time, i.e. their 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. The second reason is that mortality rates tend to decline over time, and this may happen at different rates for people in different socioeconomic groups.[3] Hence, the static picture that emerges from differences in life expectancy by different socio-economic groups does not represent anybodys actual expected longevity.

[1] See, for instance, Preston and Elo (1995) 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).

[2] See, for instance, Brown (2002) or Meara, Richards and Cutler (2008) for education, or Lin *et al.* (2003) and Singh and Siahpush (2006) for other measures of socioeconomic status.

[3] The time-changing problem of mortality rates is very well known and 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 do not need to 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.

The first contribution of this paper is to develop a new methodology for computing the expected duration of people's lives that addresses these two problems. We call this measure *expected longevity*, and we document its socioeconomic gradient. In particular, 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 and on individual health h at all ages (after age 50), and age-specific joint transition probabilities for this characteristic z and health h. 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. The transitions for the socioeconomic characteristic z allow us to address the changes in socioeconomic status over the life cycle. Using information on health h allows us to partly address the changes over time in survival rates (and in transitions for z). Our methodology borrows from the literature on unemployment duration analysis: we estimate a hazard model for survival with time-varying stochastic endogenous covariates and use it to compute expected durations.[4] However, because the expected life length after age 50 is much longer than any unemployment spell, we face a more severe right-censoring problem than that found in the labor literature. We overcome this problem by using data for individuals from different cohorts. 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. Our approach takes this direction, but uses information from different cohorts to overcome the restrictions imposed by available data.

Our results uncover a large amount of heterogeneity in expected longevities, show how important it is to keep track of the evolution of the socioeconomic characteristics z, and predict important changes in the future mortality differentials between socioeconomic groups. To see these results, we proceed in steps.

First, in order to focus on the importance of life cycle changes in socioeconomic status, we ignore information on health and, hence, assume that mortality rates and transition functions are constant over time. We find that the difference in expected longevity at age 50 between a college graduate and a high school dropout is 5.8 years for white males and 5.9 for white females. The difference between an individual at the top quintile of the wealth distribution and one at the bottom quintile is 3.1 years for males and 2.7 for females. The difference in expected longevity between individuals working full-time in the

[4] See Lancaster (1990) for an overview of duration analysis, mainly in the context of the labor market.

labor market (or unemployed looking for a job) and inactive individuals is 1.4 years for males and 0.7 for females. The difference between married and nonmarried individuals is 2.2 years for males and 1.2 for females. The difference between non-smokers and smokers is 2.2 years for males and 1.8 for females. More importantly, despite being correlated, each of these characteristics carries information on their own. When we compute these longevity differentials for different education categories, they are still large, and more so for less educated individuals. When computing the expected longevities within narrower categories, the amount of heterogeneity that we find is enormous: for instance, the expected longevity of a 50-year-old white male with a college degree who is married and does not smoke is almost 11 years greater than the expected longevity of a 50-year-old white male without a high school degree who is not married and smokes.

Our findings show the importance of keeping track of the evolution of the characteristics z that we are interested in and how unreliable the naive approach of ignoring the dynamics of characteristics can be. For instance, a male who would consistently be in the top quintile of the wealth distribution throughout his life would have an expected longevity of 10.8 more years than another male who would always be in the lowest quintile. Additionally, a male who remained married all his life would enjoy, at age 50, 4.9 more years of expected life than a nonmarried male who remained nonmarried. However, most people are not always in the same quintile of the wealth distribution or remain in the same marital status throughout their lifetimes. Our calculations take these movements into account, and as the numbers of the previous paragraph show, ignoring the dynamics in the socio-economic variables severely overestimates the socio-economic gradient in expected longevity.

In a second step, we exploit information on individual self-assessed health, which has been shown to predict mortality very well, in order to improve the prediction of the future mortality rates of the current young, hence partly overcoming the problem of time-changing mortality rates.[5] Our results show that the socioeconomic gradient of expected longevity increases when we use information on health to predict future mortality rates. For instance, the expected longevity differentials between education categories increase to 6.1 years for males and 6.0 for females (compared with 5.8 and 5.9 when ignoring information on health), and for wealth categories they increase to 3.8 years for males and 3.7 for females (compared with 3.1 and 2.7). The reason for these results is that the gap

[5] For the predictive power of self-rated health, see, for instance, Idler and Benyamini (1997) and Idler and Benyamini (1999) and references therein. Or see also our own results within this paper.

in health condition between the most and the least advantaged types within the current young cohorts today is larger than it was in the past for the current old. This implies that the socioeconomic gradient in mortality rates is likely to increase in the near future.

The second contribution of this paper, intimately linked to the first one, is to decompose the socioeconomic differences in expected longevity into differences in observable health and differences in mortality not explained by differences in observed health. In particular, we decompose the expected longevity gradients into three parts: (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. 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. Around two-thirds of them arise due to the health protection effect of education and wealth over the next years. Interestingly, the effects of education and wealth on two-year mortality rates are very small or null after controlling for self-assessed health. This finding is surprising, as both higher wealth and education suggest a higher ability to pay for terminal medical treatment. We conclude from this finding that, while financial resources may be health protecting, they are not key for mortality at the onset of terminal diseases. Instead, when looking at the decompositions of the life expectancy differentials for marital status or smoking behavior, we find that half of the differential is due to differences in 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 that we also 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. In particular, between 1992 and 2008, the expected longevity premium of college-educated individuals at age 50 has increased by 1.8 years for white males and 1.7 years for white females. The difference between individuals in the top and bottom wealth quintiles has increased by 1.4 years for white males and 0.7 years for white females. These numbers are large, and they may be easy to interpret in the context of raising income and wealth inequality during these years. However, the differential between married and nonmarried individuals has also increased for both males and females by 1.0 and 1.5 years, which may suggest that other types of explanations are also needed.

An additional advantage of our methodology is that it can be used to compute expected longevities using similar data sets in other countries, 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 mortality between different countries.[6] In particular, since the life cycle changes in socioeconomic conditions may be different in different countries, our methodology is particularly useful in correcting the biased picture that emerges from comparing mortality differentials alone.

Understanding expected longevities conditional on given socioeconomic characteristics is very important for at least three different reasons. First, any attempt to measure the welfare consequences of economic inequality should factor in the life expectancy differentials associated with it. For example, the large difference in life expectancy between college and non-college workers dwarfs the welfare impact of the income differences between these two types of individuals. Following the literature that compares welfare differences between countries, we use estimates from a measure called the *value of a statistical life* to assess the welfare differences between education groups. We find these differences to be huge: the compensated variation between a high school dropout and a college graduate is two-thirds when only consumption differences are taken into account, but increases up to 5 when we also add the differences in life expectancy. 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—is 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. And third, some financial products like life annuities, life insurance, or medical insurance are intimately related to life expectancies. Attempts to understand household portfolios of these products can potentially benefit from a deepened understanding of differences in life expectancies.

The remainder of the paper is organized as follows. In Section 2 we explain the main features of the HRS data set and our sample choices. In Section 3 we compute the socioeconomic gradients of life expectancies, which assume that people's characteristics do not change over the life cycle and that mortality rates are constant over time. Then, in Section 4 we compute the socioeconomic gradients in expected longevities, taking into account the evolution of socioeconomic characteristics over the life cycle, and we also study the evolution of these expected longevities between 1992 and 2008. In Section 5

[6] See, for instance, Majer *et al.* (2010).

we use the information on self-rated health to predict the time changes of mortality rates of different groups in the coming years and hence improve on our measure of expected longevity differentials. Then, in Section 6 we decompose the socioeconomic gradients of expected longevity. In Section 7 we provide a back-of-the-envelope measure of the welfare consequences of the large differentials in expected longevity. Finally, Section 8 concludes.

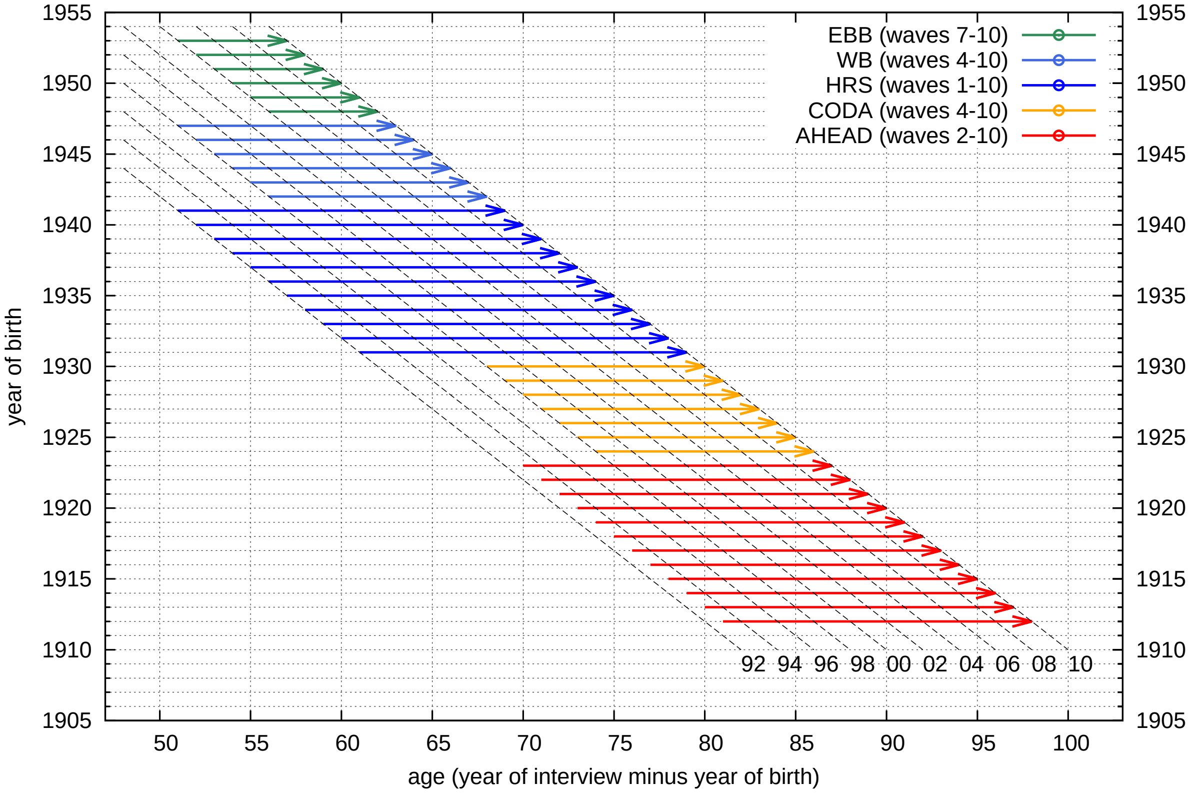
2 HRS data

The Health and Retirement Study is a biannual panel of individual level data. The panel starts in 1992, and data are now available up to the interview of 2010 (the HRS has 10 waves), which allows to compute survival rates up to the year 2008. 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 (people born before 1924), CODA (people born between 1924 and 1930), WB (people born between 1942 and 1947), and EBB (people born between 1948 and 1953). Therefore, the overall sample contains respondents age 50 or older and spouses of any age. Figure 1 provides a description of the age-cohort structure of the HRS target population. On the Y-axis we display the year of birth of an individual and on the X-axis his or her age. 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. For instance, information for 60-year-old individuals comes from those born between 1932 and 1950 who are observed between 1992 and 2010, whereas information for 70-year-old individuals comes from those born in 1923 and observed in 1993, and those born between 1928 and 1940 who are observed between 1998 and 2010. As seen in Figure 1, and due to the unbalanced entry of new cohorts, the age structure of the target population is different in every sample year. In particular, note that only in the year 2004 (seventh wave) do we have a sample of respondents of all ages in the range between 50 and 94 years.

2.1 Sample selection

To create our sample, we drop individuals for which we cannot obtain their race, sex, education, self-rated health, or survival status into the next interview. We keep observations for the age range 50 to 94. We then create two samples, one for white males and

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.

another for white females. We estimate survival rates and transition functions conditional on education, marital status, labor market status, wealth, smoking status and self-rated health. Some of these variables present missing data for a few observations. Aside from education and self-rated health, we do not drop observations with missing data, 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 56,322 individual-year observations for males and 72,516 for females, which correspond to 10,691 males and 13,288 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 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 life expectancies for gender-race subgroups.

In principle, one can use the HRS data in the year 2004 to compute life expectancies at age 50 like the NVSS does with the death rates of the total population. However, due to the relatively small sample size of the HRS, the age-specific mortality rates would be computed with very few individuals. Computing life expectancies for the overall population is not a big problem, but once we start conditioning on different observables, the size of every cell becomes very small. To overcome the sample size problem, we pool our data for all the sample years. This implies using data from individuals in several cohorts to compute each age-specific survival probability. In contrast, the use of single year age-specific survival probabilities uses individuals born in a different year to compute the death rate at every different age. Since the HRS sample period is between 1992 and 2008, we view our life expectancies as a kind of average of those reported by the NVSS between these years.[7]

3.1 Average life expectancy

We start by computing the life expectancy at age 50 for white males and for white females. We use our 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 by use of a logit model with a linear term in age. Details on this and all the remaining estimations can be found in Appendix C and Appendix D. Let us define x_a as the number of people alive at age a out of a given initial population at age 50. Then, the life

[7] See Appendix B for the evolution of the population life expectancies reported by the NVSS during these years, and Section 4.3 for our own analysis of time trends.

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 biannual 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 $A \equiv \{50, 52, ..., 94\}$. Our results are reported in the first column of Table 1.

Table 1: Life expectancies by characteristic z

	LE	Life expectancy premia					
		edu	wea	lms	mar	smok	m-s
Male	78.2	5.8	10.8	9.6	4.9	6.8	11.0
Female	81.9	5.9	9.6	7.3	3.1	5.3	7.4

Notes: The first column reports the 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), 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.

Using the whole HRS sample, we find that life expectancy at age 50 is 78.2 years for white males and 81.9 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.[8]

3.2 The socioeconomic gradient in life expectancy

The age-specific mortality rates vary substantially with variables related to socioeconomic status. In particular, it is well known that the more educated, the wealthier, or the married have lower mortality rates.[9] We can aggregate these differences in mortality rates

[8] In Appendix B we show more details of the comparison to the NVSS, and we also look at the life tables and life expectancy for the HRS in 2004 and compare them to the ones from the NVSS.

[9] See the NVSS reports for the negative relationship between mortality and education and mortality and marital status. See Attanasio and Hoynes (2000) for the negative relationship between mortality and

by computing life expectancies for each group. In particular, we want to 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, high school graduate, and high school dropout), wealth (by quintiles), labor market status (attached, semiattached, and inactive), and marital status (married and its complement). We also consider smoking behavior, which is not normally thought of as a socioeconomic characteristic but we find to be highly relevant. We also look at a four-category variable created by combining marital status and smoking. The interpretation of 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.

To compute z-specific life expectancies, we first estimate age-specific two-year survival probabilities $\gamma_a(z)$ for every $z \in Z$. To do so, we estimate 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. Then, the life expectancy $e_{50}(z_j)$ at age 50 for individuals whose z was equal to z_j at age 50 is 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. We find very large socioeconomic gradients of life expectancies, and less so for females (except in the case of education). At age 50, the difference in life expectancy between a college graduate and a high school dropout is 5.8 years for males and 5.9 for females. The difference between an individual at the top quintile of the wealth distribution and one at the bottom quintile is 10.8 years for males and 9.6 for females. The difference between individuals strongly attached to the labor force and inactive individuals is 9.6 for males and 7.3 for females. The difference between married and nonmarried individuals is 4.9 years for males and 3.1 for females. The difference between nonsmokers and smokers is 6.8 years for males and 5.3 for females. Combining marital status and smoking also shows

wealth.

large differentials: the difference between married nonsmokers and nonmarried smokers is 11.0 years for males and 7.4 for females.

Of course, these results are silent about causality. It may be that college, wealth, or marriage help to ensure survival. Or it may be that people differ in some unobserved characteristics and that individuals with lower mortality rates self-select into studying more, accumulating more wealth, and getting and staying married.[10] Since people die only once, there is no possibility of including fixed effects in the estimation. Alternatively, one could try to add random effects. We do not want to do this, however, because identification could come only from functional form assumptions. In addition, there is also the problem of interpretation once we move into estimating the transition matrices of socioeconomic characteristics: it is not clear how to identify the unobserved heterogeneity in both the hazard rate and the law of motion of the covariates at the same time.

4 Expected longevities

When the characteristic of interest z changes over the life cycle, even if the age-dependent mortality rates by type z do not change over time, the difference in life expectancy between groups is not a good proxy for the difference in expected longevity. The reason is that the hypothetical group of age 50 individuals with characteristics $z = z_j$ expect z to change over the years and hence later in life may well face the mortality rates of $z \neq z_j$. This is not a problem with education, which is fixed at age 50. But wealth, marital status, and smoking behavior change substantially over the years. Even more important are the changes in labor market status because people clearly drop from the labor force as they age. For this reason, we want to focus on a way of aggregating socioeconomic differences in mortality that takes into account the dynamic nature of socioeconomic status.

Accordingly, we compute 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. To compute these statistics, we need two elements: 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 us define $x_a(z)$ as the fraction of people who are alive and of type z at age a out of a given population at age 50. Then,

[10] For instance, Heckman (2011) documents a relationship between conscientiousness and mortality on the one hand and conscientiousness and socioeconomic status on the other.

expected longevity $\ell_{50}(z_j)$ at age 50 conditional on $z = z_j$ at age 50 is computed as

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

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 $p_a(z'|z)$. Note that at this point, we are ignoring the possible effects of health on γ_a and p_a, something that we will address in Section 5.

Table 2: Longevity differentials by characteristic z

	edu	wea	lms	mar	smok	m-s
Male						
(1) All	5.8	3.1	1.4	2.2	2.2	5.1
(2) College graduates	–	2.3	1.6	1.7	2.3	3.3
(3) High school dropouts	–	2.6	2.0	2.6	2.3	5.6
(4) Education and z	–	7.3	7.3	8.3	7.8	10.7
Female						
(1) All	5.9	2.7	0.7	1.2	1.8	2.9
(2) College graduates	–	0.7	0.5	1.5	1.6	2.1
(3) High school dropouts	–	1.9	0.7	1.9	1.4	2.8
(4) Education and z	–	6.8	6.3	7.8	7.0	8.4

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 high school dropouts. Rows (4) report the difference between individuals with a college degree and the most advantaged type z and individuals that are high school dropouts and of the least advantaged type z.

4.1 The socioeconomic gradient in expected longevity

In rows (1) of Table 2, we report the differentials in expected longevity for the same groups as in Table 1. Of course, there is no difference for education because education does not change over the life cycle. For the rest of the cases, we still find a very important amount of heterogeneity in expected longevities at age 50, but less than in life expectancies. At age 50, the difference between an individual at the top quintile of the wealth distribution and one at the bottom quintile is 3.1 years for males and 2.7 for females. The difference between individuals strongly attached to the labor force and inactive individuals is 1.4 for males and 0.7 for females. The difference between married and nonmarried individuals is 2.2 years for males and 1.2 for females. The difference between non-smokers and smokers is 2.2 years for males and 1.8 for females. The difference between married nonsmokers and nonmarried smokers is 5.1 years for males and 2.9 for females.

As expected, the changing nature of the socioeconomic characteristics is important in order to accurately measure the differentials in expected longevity between different socioeconomic groups. The concept of life expectancy is a measure of expected longevity that imposes an identity matrix for the transition $p_a(z'|z)$ of characteristic z. The differences between Tables 1 and 2 show that ignoring the changing nature of socioeconomic characteristics leads to very important biases in our measures of longevity differentials. Previous measurements of life expectancy differentials ignored the transitions, and hence they overstated the socioeconomic gradient of expected longevities. See, for instance, Lin *et al.* (2003)—who compute life expectancies based on marital status, labor market status or family income by using the National Longitudinal Mortality Study (NLMS)—or Singh and Siahpush (2006)—who use aggregates of county-level variables.

4.2 The socioeconomic gradient within education groups

We also want to explore to what extent the differences in longevity associated with different socioeconomic variables are independent from each other. In particular, more educated people tend to be richer, divorce less, smoke less, or be more active in the labor market at older ages. For this reason, 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 high school dropouts. We find that the differentials remain large within education groups, for both men and women. Furthermore, for wealth, labor force status and marital status, we observe larger differentials within high school dropouts than within college graduates.

Finally, in row (4) of Table 2, we report the longevity differential between college graduates with the most advantaged type z and high school dropouts with the least advantaged type. This measure shows very large differentials, larger than the average differential by education group, which confirms the importance of characteristic z beyond education. For instance, the longevity differential between a male in the top wealth quintile and a college degree and a male in the bottom wealth quintile and no high school degree is 7.3 years, whereas the average education differential is only 5.8 years and the average wealth differential is only 3.1.

4.3 Time trends

The data from NVSS show an upward trend in life expectancies between 1992 and 2008. Life expectancies at age 50 have increased by 2.2 years for white males and by 1.0 years for white females. There is no reason to expect that these increases have been homogeneous across socioeconomic groups. Indeed, Meara *et al.* (2008) report an increase in the educational differential of life expectancies between 1990 and 2000. 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. In particular, we add a linear year term independent of age but z-type dependent. By doing so, 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.[11] With these estimates, we can compute expected longevities as in Table 2 but specific to every year in our sample. We have to interpret these year-by-year expected longevities as the ones that would arise if individuals had to face 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 expected longevity differentials for the first and last years of our sample period, as well as the change over these 18 years. Our estimates show an increase of 1.6 years in the life expectancy at age 50 for white males between 1992 and 2008. The corresponding increase reported by the NVSS is 2.2 years. Hence, our HRS sample captures the trend but misses 0.6 years. By contrast, the life expectancy for white females falls by half a year in our sample, whereas it increases by 1.0 years in the NVSS data.

Regarding the longevity differentials, we find that they have all been increasing over time with no exception. In particular, the educational differential for white males in-

[11] 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 (1992) for details.

Table 3: Time trends

	LE	Longevity differentials					
		edu	wea	lms	mar	smok	m-s
Male							
1992	77.4	4.7	2.3	1.0	1.6	1.6	3.7
2008	79.0	6.5	3.7	1.7	2.6	2.8	6.2
Δ	+1.6	+1.8	+1.4	+0.7	+1.0	+1.2	+2.5
Δ_{NVSS}	+2.2						
Female							
1992	82.2	5.1	2.3	0.3	0.3	1.4	1.4
2008	81.7	6.8	3.0	0.9	1.8	2.3	4.2
Δ	-0.5	+1.7	+0.7	+0.6	+1.5	+0.9	+2.8
Δ_{NVSS}	+1.0						

Notes: The first column reports the 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 in our HRS data set. Δ_{NVSS} refers to the same difference according to the NVSS data.

creased by 1.8 years, from 4.7 years in 1992 to 6.5 years in 2008. For white females it increased by 1.7 years, from 5.1 to 6.8. Preston and Elo (1995) already showed evidence of an increase in the education gradient of mortality rates between 1960 and the early 1980s.[12] Meara *et al.* (2008) show an increase in the life expectancy gradient of education between the 1980s and the 1990s, and also between 1990 and 2000.[13] For the period between 1990 and 2000, they report that the education differential of life expectancy at age 25 increases by 1.6 years for white males and 1.9 for white females. Our results for education show that this trend has extended into the period 1992-2008. 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 has increased, and so has wealth inequality.[14] 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

[12] 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.

[13] 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.

[14] See Heathcote *et al.* (2010) for the increase in labor income inequality and the college premium, and Díaz-Giménez *et al.* (1997) and Díaz-Giménez *et al.* (2011) for the increase in wealth inequality.

may also be due to an increase in the selection of long-lived individuals into marriage and nonsmoking.

4.4 Summary

We conclude that life expectancies are an ineffective way of aggregating the socioeconomic differences in mortality rates when the socioeconomic characteristics of interest may change over the life cycle; that different socioeconomic variables carry independent relevant information to predict longevity differentials (in particular wealth, labor market status, or marital status play a very important role independently of education); that there is more heterogeneity within males than within females; and that there is somewhat more heterogeneity within less educated than within more educated individuals; and finally, that these differentials in expected longevity between socioeonomic groups have been increasing substantially during the last 17 years.

5 Expected longevities with information about health

One advantage of the HRS is that it provides several measures of the health status of individuals. Chief among them is the information on self-assessed health. Self-assessed health is a subjective measure that comes from the respondent's answer when asked to evaluate her general health level. This answer can be one of five categories: excellent, very good, good, fair, and poor. As is well known, self-assessed health is a very important determinant of survival, even after controlling for socioeconomic characteristics and measured health conditions.[15] In addition, self-rated health is a very interesting measure of health because it is present in several data sets of household survey data commonly used by economists, such as the Panel Study of Income Dynamics (PSID) and the National Longitudinal Study of Youth (NLSY), as well as the HRS. Partly for this reason, there is a growing quantitative literature that uses heterogeneity in self-rated health to predict heterogeneity in mortality risk.[16] For these reasons, in order to improve on our measures of expected longevity, we will use self-assessed health as a summary of the stock of health of individuals.

[15] See, for instance, Idler and Benyamini (1997) and Idler and Benyamini (1999) and references therein.

[16] See, for instance, De Nardi *et al.* (2010), Kopecky and Koreshkova (2011), and Nakajima and Telyukova (2011).

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 in the data an age-dependent survival function $\gamma_a(h)$ as a function of health only, an age-dependent health transition function $p_a(h'|h)$, and the initial health distribution $\varphi_{50}(h)$. Again, we estimate logistic and multinomial logistic regressions 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 compute the expected longevity ℓ_{50}^{h} as

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

The statistic ℓ_{50}^{h} shall be interpreted as 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, but may differ on the initial distribution of health $\varphi_{50}(h)$. Hence, compared with the *period* life expectancy e_{50} of Section 3, our measure ℓ_{50}^{h} takes into account that the 50-year-olds of today may face in the future mortality rates γ_a that are different from those faced by the current old. The measure 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. Admittedly, it ignores possible future changes in $\gamma_a(h)$ and $p_a(h'|h)$, something for which we do not have any information. Hence, if the health distribution at age 50 of those cohorts that are age 50 today is better than it was for the older ones when they were 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 78.4 years for white males and 82.1 years for white females. These values are very similar to the life expectancies e_{50} computed in Section 3 (see Table 1, column 1), which were 78.2 and 81.9, respectively. This indicates very minor differences in initial health at age 50 in favor of the current young cohorts. Hence, if the trend gains in life expectancy over the last years are going to extend into the future, it will not be through the better health of the 50-year-olds, but rather through improvements over time of $\gamma_a(h)$ and $p_a(h'|h)$.

Table 4: Expected longevities with health status

	EL	Expected longevity differentials					
		edu	wea	lms	mar	smok	m-s
Male							
All type-specific	78.4	6.1	3.8	3.4	2.5	2.9	6.1
(a) type-specific initial health		1.7	1.3	2.1	0.4	0.5	1.0
(b) type-specific transition		4.7	1.7	0.6	0.7	1.0	1.7
(c) type-specific mortality		0.0	0.9	0.4	1.4	1.5	3.5
Female							
All type-specific	82.1	6.0	3.7	1.3	1.4	2.3	3.5
(a) type-specific initial health		1.1	1.2	0.8	0.3	0.2	0.4
(b) type-specific transition		4.9	1.7	0.3	0.7	0.9	1.5
(c) type-specific mortality		0.3	0.8	0.2	0.4	1.2	1.7

Note: See column labels in footnote of Table 1.

5.2 The socioeconomic gradient in expected longevity

Next, we use the information on health in order to improve on our measures of the socioeconomic gradient of expected longevity computed in Section 4.1. We will need three different objects: First, the health distribution at age 50 for every type z, $\varphi_{50}(h|z)$; second, the age-dependent joint health and characteristic z transition matrix, $p_a(z', h'|z, h)$; and third, 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, but they are upgraded to include dummies for each element in $Z \times H$ and their interaction with age.[17] Let us define $x_a(z, h)$ as 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

[17] 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). We prefer our specification for two reasons. First, our specification 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, our specification 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 C and D). The decompositions in Section 6 are based precisely on this distinction.

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 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 2-7 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 6.1 years when using the information on health (whereas it is only 5.8 years when not using it); for wealth it is 3.8 years (3.1); for labor market status it is 3.4 years (1.4); for marital status it is 2.5 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: today, the least advantaged types start at age 50 with worse health relative to the best types than they did in the past. In particular, the health distribution among the current old high school dropouts, wealth poor, individuals detached from the labor market, nonmarried and smokers is better than it will be in the future when the current young cohorts age. Hence, we should expect the socioeconomic gradient in mortality rates to increase in the coming years.

5.3 Summary

The use of information on health status to predict future aggregate mortality rates does not improve on the use of raw life expectancies; hence, the information on health status does not predict an improvement of mortality rates in the near future. If there is any improvement in life expectancies in the coming years, it will come from reductions of mortality conditional on health. Instead, when we use the information on health to predict future mortality rates of different socioeconomic groups and to predict the future transition matrices of the socioeconomic characteristics, then the obtained differences in the socioeconomic gradient of expected longevities increase; hence, we expect the differences in mortality rates across socioeconomic groups to increase in the coming years.

6 Decomposing the socioeconomic gradient in expected longevity

In the previous section we exploited information about health in order to help predict future mortality rates and future transition matrices for the current young. In this section we exploit this same information about health to decompose the expected longevity gradients into three parts: (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 $\{\varphi_{50}(h), p_a(h'|h), \gamma_a(h)\}$. We show these results in Table 4. When we look at education, we find that around one-third 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-educated individuals report better self-rated health than high school dropouts. 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 1.7 years of life expectancy difference for males and 1.1 for females. The education-specific health transition matrix, $p_a(z', h'|z, h)$, accounts for two-thirds of the education gap (see rows (b) for both males and females). That is, if health at age 50 were equally distributed for college and high school dropouts and mortality rates were dependent on health but not on education, the fact that self-rated health deteriorates less for highly educated individuals generates by itself a life expectancy gap of 4.7 years for males and 4.9 for females. Finally, the effect of education-specific mortality rates $\gamma_a(z,h)$ is very small: 0.0 years for males and 0.3 years for females (see rows (c)). Indeed, in the underlying logit regressions, the effect of education on mortality rates for males is statistically not different from zero once we control for health; see Appendix C for details. The decomposition for the life expectancy gaps by wealth generates a similar pattern, in which initial health differences account for roughly one-third and the health-protecting nature of wealth accounts for around two-thirds, whereas mortality rates are less influenced by wealth.

Instead, the decomposition results are substantially different for smoking and marital status. Again, type differences in initial health and health transitions matter. However, mortality rates are smaller for married or nonsmokers even after controlling for self-rated health. We find that the same is true when we look at both types together. In particular, type-specific mortality rates account for around half 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 2.1 years out of 3.4 for males and 0.8 out of 1.3 for females. This is consistent with the evidence that early retirement / inactivity is very much linked to health status.

It is interesting to note the difference between education and wealth on the one side, and marital status and smoking on the other. The education and wealth results suggest that financial resources may be health protecting, but they are not key for mortality conditional on health. Hence, once health has fallen with some terminal condition, financial resources do not help much. Instead, when looking at the decompositions of the life expectancy differentials for marital status or smoking behavior, we find that half of the differential is due to differences in mortality rates even after controlling for differences in measured health. This finding is very interesting because it shows some advantage of being married or not smoking that is not necessarily related to access to more financial resources.

7 Welfare implications of life expectancy heterogeneity

In this paper we have been documenting the size of properly measured differences in expected longevities between different groups at age 50. To illustrate the importance of these differences, we want to compute the compensated variation between education types. Consider the fact that the consumption of college graduates is about two-thirds higher than that of high school dropouts (Attanasio *et al.* (2011)). Imagine that both groups had the same life expectancy, the same rate of consumption growth, and the same constant relative risk aversion (CRRA) utility function. The compensated variation of the types would then be exactly two-thirds. That is, a high school dropout would be indifferent between increasing her consumption by two-thirds or getting the consumption bundle of a college graduate.

However, expected longevity for college graduates is 6.1 years greater than that for high school dropouts (see Table 4). We want to measure how much greater should be the compensated variation of the types due to this difference in expected longevities. This exercise is analogous to those done to compare cross-country differences in welfare (and their evolution) when considering life expectancy differences as well as GDP differences.[18] To do so, we use a purposely simple framework and abstract from life-cycle considerations by assuming a constant consumption flow and a constant survival probability for each type. Consider a 50-year-old with a consumption flow c and a survival probability γ. Her value function is then given by

$$V\left(c,\gamma\right)=\sum_{a=50}^{\infty}\left(\beta\gamma\right)^{a-50}u\left(c\right)=u\left(c\right)\frac{1}{1-\beta\gamma},$$

where β is the subjective time discount factor and $u\left(c\right)$ is the period utility function. Now, what is the equivalent compensation that would make a high school dropout (HSD) indifferent to being a college graduate (CG)? This is the quantity Δ that solves

$$V\big(c_{HSD}\left(1+\Delta\right),\gamma_{HSD}\big)=V\big(c_{CG},\gamma_{CG}\big).$$

Hence, Δ is implicitly defined by

$$\frac{u\left(c_{HSD}\left(1+\Delta\right)\right)}{u\left(c_{CG}\right)}=\frac{1-\beta\gamma_{HSD}}{1-\beta\gamma_{CG}}.$$

Of course, with $\gamma_{HSD}=\gamma_{CG}$ and CRRA utility, we have $\Delta=c_{CG}/c_{HSD}-1=2/3$. Now, with differences in expected longevity between types, we need a utility function that gives some value to life. We use the standard CRRA specification with an added constant α:

$$u\left(c\right)=\alpha+\frac{c^{1-\sigma}-1}{1-\sigma}$$

[18] See, for instance, Becker, Philipson and Soares (2005) and Jones and Klenow (2010).

To compute Δ we need to give values to the parameters α, σ, β, γ_{CG}, and γ_{HSD}, as well as to the consumption flows c_{CG} and c_{HSD}. Let us take the standard values $\beta = 0.96$ and $\sigma = 1$. For the expected longevities, we have computed that a 50-year-old college graduate expects to live 31.5 more years and a high school dropout 25.3 years. This recovers constant survival rates $\gamma(z_{CG}) = 0.968$ and $\gamma(z_{HSD}) = 0.960$. The most difficult parameter to pin down is the value of life α. We are going to use the *value of a statistical life* (VSL) measure that emerges from the literature on wage compensating differentials for dangerous jobs. Kniesner *et al.* (2012), exploiting the panel dimension of the PSID, argue that the average VSL in 2000 is between \$4 and \$10 million. Hall and Jones (2007) use a value of \$3 million and Jones and Klenow (2010) \$4 million. To stay on the conservative side, we take the lower end of the range, \$4 million. Now, in order to use the VSL to recover α, a we follow the standard approach of Murphy and Topel (2003). Note that by setting

$$V_c(c,\gamma)\,dc + V_\gamma(c,\gamma)\,d\gamma = 0,$$

we can obtain the model's implied value of life:

$$\frac{dc}{d\gamma} = -\frac{u(c)}{u'(c)}\left(\frac{\beta}{1-\beta\gamma}\right).$$

To recover α from the average VSL we use $\gamma = 0.965$, which generates the average expected longevity of 78.3 (see Table 4) and $c = \$24,187$, which was the average consumption per capita in the U.S. economy in 2000. We obtain $\alpha = 2.64$. Finally, we use $c_{CG} = \$27,684$ and $c_{HSD} = \$16,610$, which is consistent with college graduates enjoying a consumption flow two-thirds higher than high school dropouts and also with the average consumption per capita $c = \$24,187$.[19]

Given all this, we obtain $\Delta = 5.52$. This is a huge number. Note that the compensated variation ignoring life expectancy differences was only 0.67, and that by capitalizing this compensated variation to age 50 using the 6.1 year difference in life expectancy and $\beta = 0.96$, we only increase it to 0.84. Hence, properly accounting for the value in life expectancy between different socioeconomic types dwarfs any available measure of welfare differences between them.

[19] We obtain these numbers by using the proportions of 50-year-old college graduates, high school graduates, and high school dropouts in our HRS sample, and by setting the consumption for the high school graduates equal to the average.

8 Conclusions

We have developed a new methodology to compute differences in the expected longevity of individuals that have different socioeconomic characteristics—education, wealth, labor market attachment, marital status, and smoking status—at age 50. Our measure deals with the 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, that estimates a hazard model for survival with time varying stochastic endogenous covariates, is related to the literature in unemployment duration (Lancaster (1990)). We have uncovered an enormous amount of heterogeneity in expected longevities despite our corrections, for instance, a 50-year-old white male who is married, does not smoke, and has a college degree lives almost 11 more years than a 50-year-old white male who is unmarried, smokes, and did not finish high school. While our study only covers the United States, our methodology is also applicable to various European countries that have produced data sets with similar information to the one that we have used.

This large amount of heterogeneity in expected life duration matters. First, the socio-economic gradient in expected longevity dwarfs the welfare implications of the income differences accruing to different socio-economic groups. We have shown this result for education. Second, as discussed in the Introduction, 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. Therefore, any quantitative analysis of reforms of these types of policies should consider the heterogeneous expected longevities of different population subgroups. And third, 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 understand whether the pricing of these products is actuarially fair, which is itself important to understand the take up of these products.

We have also decomposed the differences in longevity into a fraction that is due to differences in self perceived health at age 50, differences in mortality among groups with different socioeconomic characteristics for each health category, and differences in the preservation of health, and have found that far and large the most important component is the advantage that various socioeconomic groups 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 opens an important questions for further research in trying to understand where does the survival advantage of married people come from.

We have also documented 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 socio-economic differences in mortality are likely to widen in the coming years.

As a final remark, the transitions of health and socioeconomic characteristics that we have constructed have other uses beyond the calculation of expected longevity. Structural models can take advantage of our transitions to tease out how much mortality differences are the result of intrinsic advantages of certain groups and how much are the result of differential investments to preserve health.

A Data

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

A.1 Variable definitions

Education. The variable RAEDUC provides five educational categories: high school dropout, 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 and hence reduce education to three categories.

Wealth. The HRS provides several measures of assets and liabilities. We define our wealth variable as total household net worth per adult, excluding second residences and mortgages on second residences. The reason for these exclusions is 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, we define total wealth as the sum of HwASTC, HwACHCK, HwACD, HwABOND, HwAOTH, HwAHOUS, HwARLES, HwATRANS, HwABSNS, and HwAIRA, minus HwDEBT, HwMORT, 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.

Labor market status. The variable RwLBRF provides seven categories for the relationship of the respondent with the labor market. We reduce it to three categories: attached, semiattached 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. We use the variable RwMSTAT and classify as married those who answer to be either married or partnered, and classify as nonmarried the rest, which includes separated, divorced, widowed, and never married.

Smoking. We use the variable RwSMOKEN that reports whether the respondent is currently a smoker.

Health. We use the 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 qthe 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, a 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'.

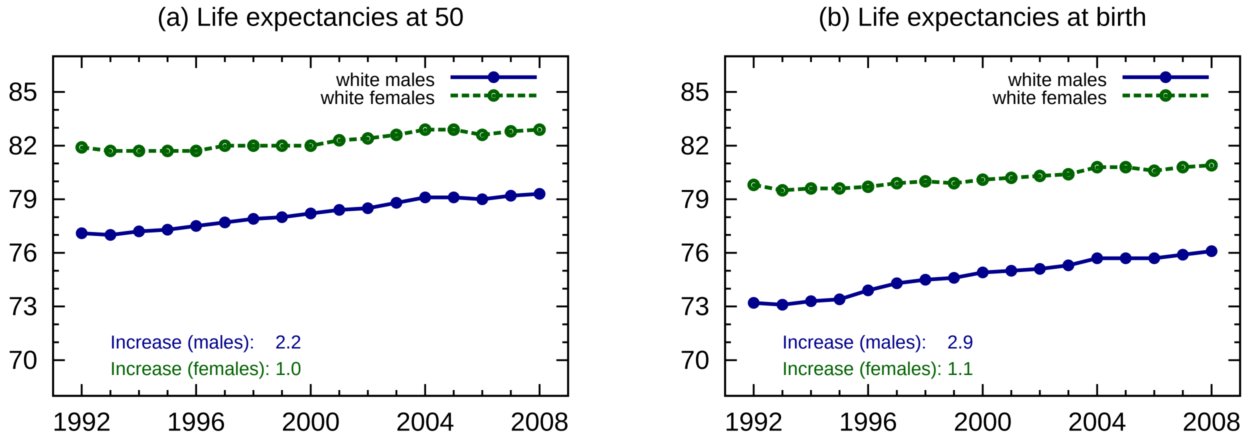
A.2 Sample selection

To select our sample, 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. This gives us 13,215 males and 17,034 females. We drop individuals with missing information for race (11 males and 8 females), drop nonwhite individuals (2,422 males and 3,511 females), and drop individuals with missing information for education (11 males and 6 females), which leaves the sample as 10,799 males and 13,517 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 62,092 and 79,406 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 (612 and 761 individual-year observations). This happens for some observations of people who could not be followed upon withdrawing from the sample (code 7 above). Next, we drop individual-year observations with missing information on health (5,158 for males and 6,129 for females, which correspond to 274 and 321 different individuals). What is left is our working sample, with 10,691 males that provide 56,322 individual-year observations and 13,288 females that provide 72,516 individual-year observations. Out of these individual-year observations, we have 3,846 deaths for males and 4,006 for females, with average death rates of 6.8% and 5.5%.

B Comparing the HRS and NVSS mortality rates

Life expectancies in the United States improved between 1992 and 2008, which is our sample period. In particular, life expectancy at 50 increased by 2.2 years for white males—from 77.1 to 79.3—and by 1.0 years for white females—from 81.9 to 82.9—(see Figure 2). This shows a reduction of the life expectancy gap between males and females of more than 1 year over this 17-year period. The life expectancies that we obtain from the HRS using all the sample years are 78.2 and 81.9 (see Table 1 in the main text). They stay within the range of the ones reported by the NVSS. The life expectancies computed by the NVSS decreased slightly in 1993 for both sexes, so their actual ranges are 77.0 to 79.3 for white males and 81.7 to 82.9 for white females.

Figure 2: Life expectancies in the United States (NVSS)

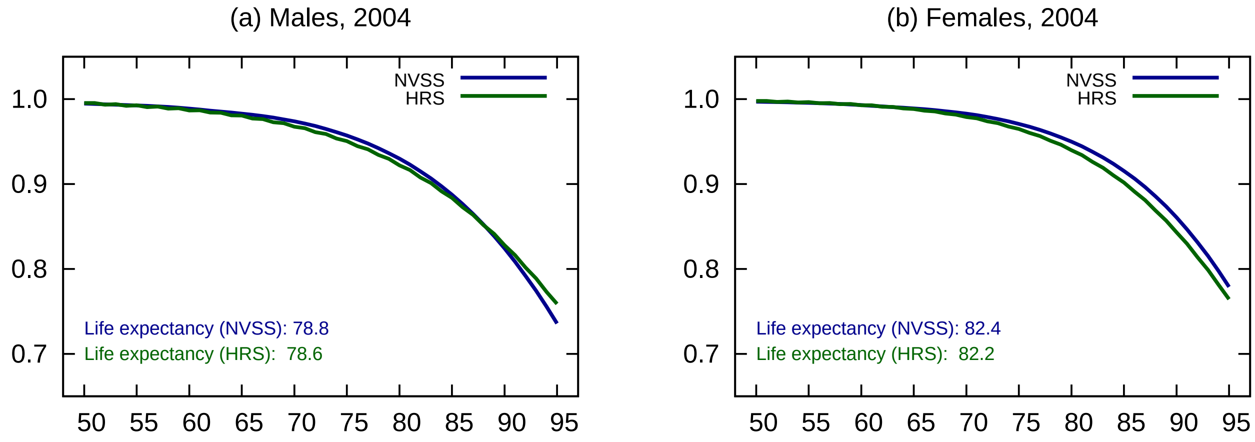


(a) Life expectancies at 50

(b) Life expectancies at birth

A more accurate comparison between the NVSS and the HRS is to use only HRS data from the year 2004, where we have data on individuals at all ages, from 50 onward. In Figure 3 we report the age-specific survival probabilities computed in the HRS and plot them alongside the ones from the NVSS for 2004.[20] The survival probabilities from the HRS are very close to the ones reported by the NVSS. The life expectancies that arise in the HRS are 78.6 for white males and 82.2 for white females, whereas the ones that arise from the NVSS data are 78.8 and 82.4. The differences are only 0.2 years in both cases.

Figure 3: Survival rates: NVSS vs. HRS (2004)



C Estimation of 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

[20] The HRS survival probabilities are two-year probabilities, but for this picture we recover the one-year probabilities by using the identity $\gamma_{a,a+2} = \gamma_{a,a+1}\gamma_{a+1,a+2}$ and using the assumption that $\gamma_{50,51} = \gamma_{51,52}$.

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 4 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 high school dropouts 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.

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 11 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 $\gamma_a(h)$ and $\gamma_a(h, z)$, where $h \in H$ is

Table 5: Logistic regressions for survival, white males

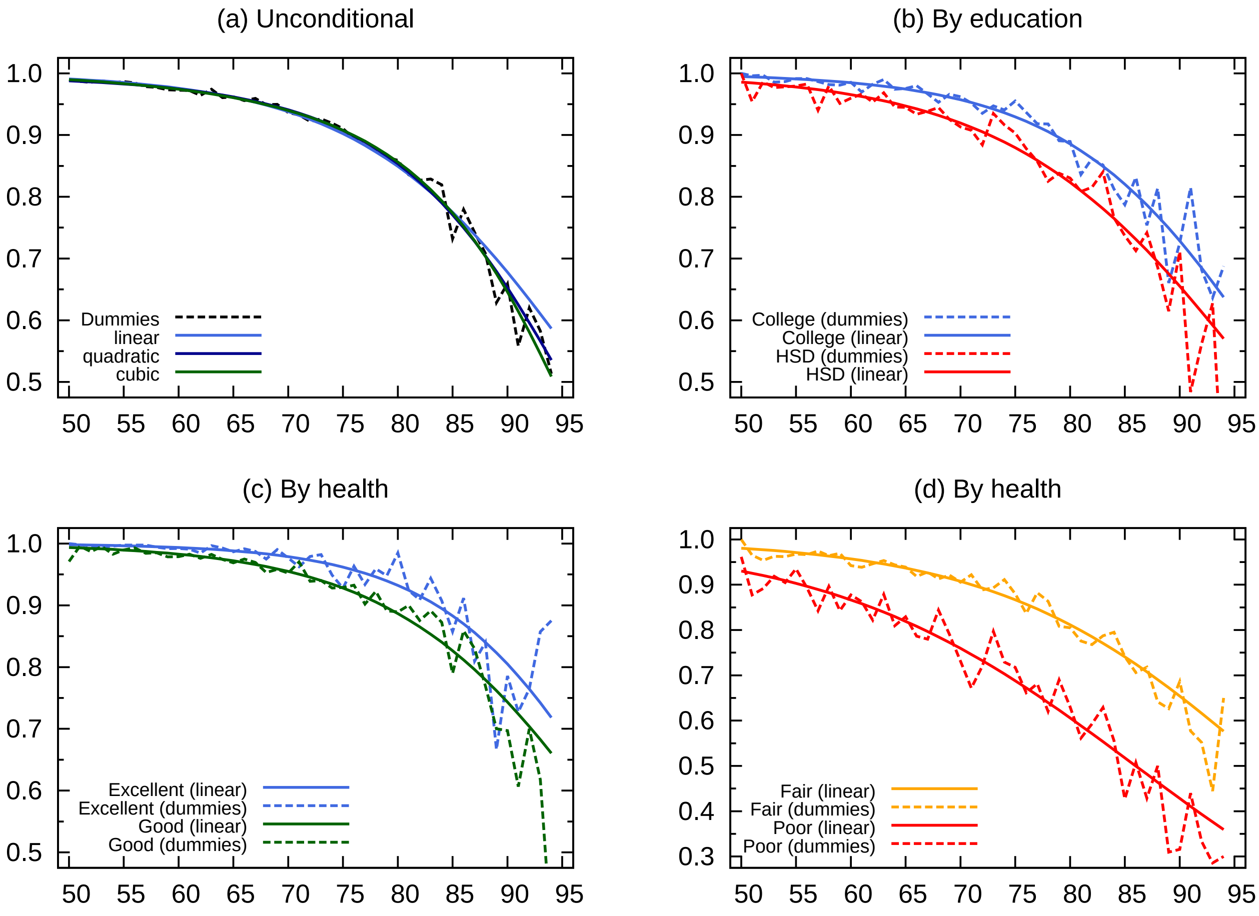
	edu	wea	lms	mar	smok	m-s
_cons	10.53***	10.89***	9.015***	9.755***	10.48***	10.46***
	(31.28)	(30.91)	(17.62)	(59.56)	(65.12)	(55.31)
$D_{z_t=z_2}$	-1.075***	-0.246	0.252	-1.565***	-1.954***	-1.413***
	(-2.78)	(-0.51)	(0.31)	(-5.22)	(-5.50)	(-3.17)
$D_{z_t=z_3}$	-1.810***	-1.134**	-1.269**			-1.037***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

Notes: t statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The omitted dummy variable corresponds to the most advantaged category, and the rest of the dummies are ordered toward the least advantaged categories.

self-rated health. We use the same logistic regression upgraded to include health:

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

Figure 4: Survival rates: parametric vs non-parametric age



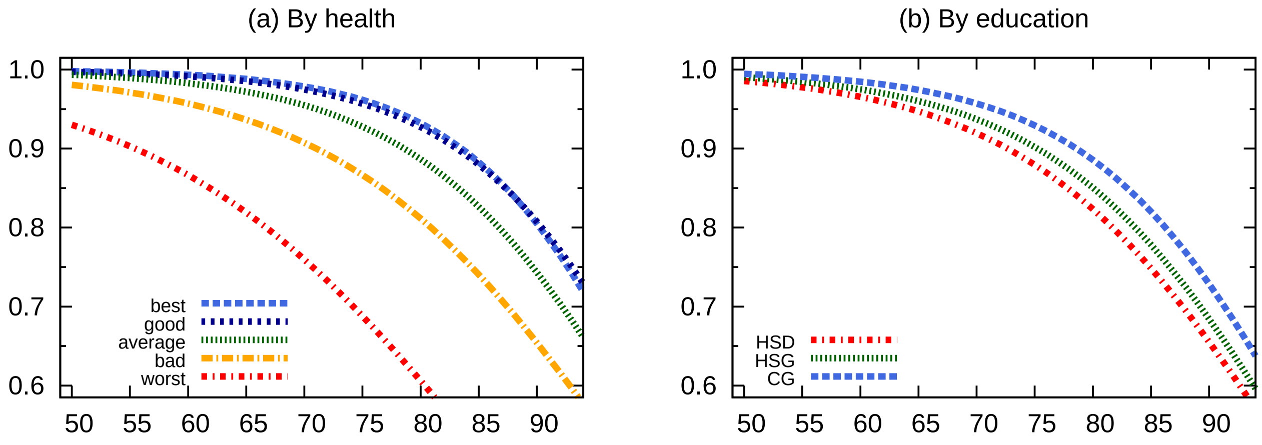
Notes: Predicted yearly survival rates, sample of white males. Since estimates correspond to two-year survivals, we report the squared root of the predictions from our logit regressions.

and

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

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 5 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 5: Survival rates: by education and by health



Notes: Predicted yearly survival rates, sample of white males. Since estimates correspond to two-year survivals, we report the squared root of the predictions from our logit regressions.

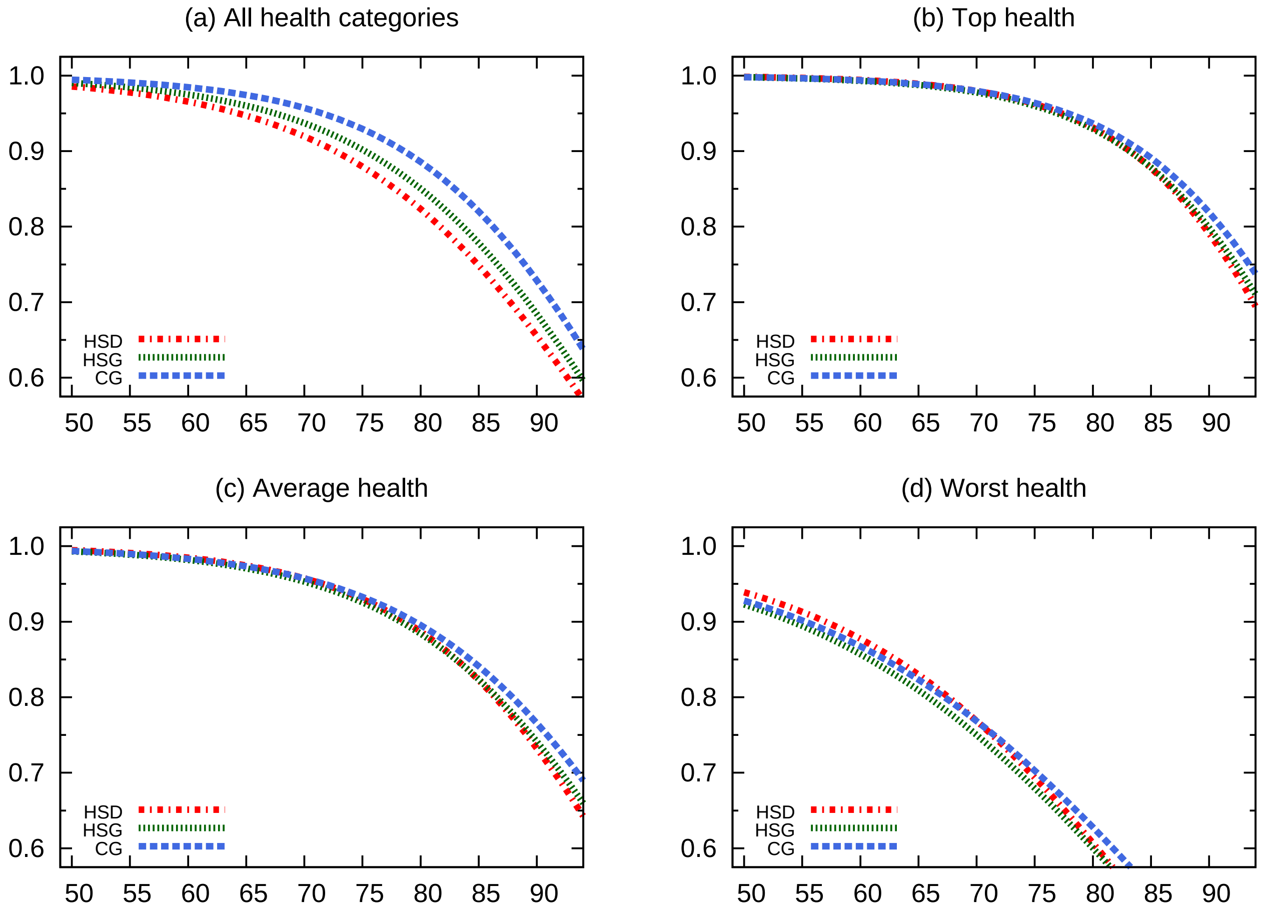
When we put both types of variables together in the same regression, we can use differences in the likelihood to test how much information education adds to health and how much information health adds to education. In Table 6 we report some results from the three logistic regressions: only health variables, 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 5, 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 very close to one and 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%.

A visual inspection of these results comes from Figure 6. In panel (a) we reproduce the survival rates by education group as in Figure 5, 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.

Table 6: Survival regressions

Logit regressions for survival (white males)						
	Odds ratios				**LR test**	
	cg vs hsd		**h2 vs h4**		χ^2	**p-value**
	65	**75**	**65**	**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		

Figure 6: Survival rates: by education and health jointly



Notes: Predicted yearly survival rates, sample of white males. Since estimates correspond to two-year survivals, we report the squared root of the predictions from our logit regressions.

D Estimation of 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 $p_{a,t}\left(z'|z\right)$, 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 $p_a\left(h'|h\right)$ and $p_a\left(z',h'|z,h\right)$, 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.

References

Attanasio, O., Battistin, E. and Padula, M. (2011). *Inequality in living standards since 1980: income tells only a small part of the story*. AEI Press, Washington, D.C.

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

Becker, G. S., Philipson, T. and Soares, R. (2005). The quantity and quality of life and the evolution of world inequality. *American Economic Review*, **95** (1), 277–291.

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.

De Nardi, M., French, E. and Jones, J. (2010). Why do the elderly save? the role of medical expenses. *Journal of Political Economy*, **118** (1), 39–75.

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.

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.

Hall, R. and Jones, C. (2007). The value of life and the rise in health spending. *Quarterly Journal of Economics*, **122** (1), 39–72.

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.

Heckman, J. J. (2011). Integrating personality psychology into economics, nBER Working Paper 17378.

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.

Jones, C. and Klenow, P. (2010). Beyond gdp? welfare across countries and time, nBER Working Paper 16352.

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

Kniesner, T., Viscusi, W., Woock, C. and Ziliak, J. (2012). The value of a statistical life: Evidence from panel data. *Review of Economics and Statistics*, **94** (1), 74–87.

Kopecky, K. and Koreshkova, T. (2011). The impact of medical and nursing home expenses on savings and welfares, mimeo.

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.

Murphy, K. and Topel, R. (2003). The economic value of medical research. In K. Murphy and R. Topel (eds.), *Measuring the Gains from Medical Research: an Economic Approach*, Chicago: University of Chicago Press.

Nakajima, M. and Telyukova, I. (2011). Home equity in retirement, mimeo.

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.

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

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