Error in the Measurement of Mortality: An Application to the Analysis of Racial Mortality Disparity

Yu-Chieh Hsu and Lowell J. Taylor
Heinz College, Carnegie Mellon University

February 21, 2011

Abstract

A large empirical literature studies the forces that shape racial disparity in mortality. Given that factors early in one's life can be important for subsequent mortality outcomes, such research often relies on panel data. An important example is the National Longitudinal Survey of Older Men (NLS-OM), which collected data for men aged 45–59 in 1966 and several subsequent years, and then also reported deaths as indicated by death certificate data collected in 1990. An important methodological issue arises in studies that use such data: Deaths are likely to be under-reported, most likely in systematic ways. In the NLS-OM, for example, the matching procedure appears to have missed a substantial number of deaths. We work out a simple model that illustrates the effect of this measurement error, and then show that inappropriate handling of the measurement error in survival analysis causes serious problems for inference.

Keywords: mortality, racial mortality disparity, measurement error

PLEASE NOTE: THIS WORK IS PRELIMINARY. We appreciate helpful comments from Daniel Nagin, Shamena Anwar, Dan Black, and Seth Sanders.

1 Introduction

A large literature demonstrates that in the United States there are very large differences in mortality rates of black and white individuals. For example, Harper, *et al.* (2007) note that while life expectancy at birth converged for blacks and whites during the period from 1900 to 1940, that gap remained large, and failed to decline consistently after the 1960s. Levine, *et al.* (2001) find that from 1979 to 1998 the "black:white ratio of age-adjusted, gender-specific mortality increased for all but one of nine causes of death that accounted for 83.4% of all US mortality in 1998."

While the proximate medical causes for the black-white gap are reasonably well known, the underlying mechanisms are not. There is ample evidence suggesting that persons of lower socioeconomic status have reduced life expectancies, but some evidence indicates that economic disparities are not the sole source of the black-white gap in mortality. For example, Sorlie, *et al.* (1992) find that increased income lowers mortality rates for everyone, but that blacks have higher mortality than whites at every level of income. Guralnik, *et al.* (1993) suggest that educational attainment may have a stronger effect than race *per se* on life expectancy.[1]

Sorting out the complicated roles of race and socioeconomic status is made all the more difficult by the fact that conditions early in one's life are likely to influence mortality later in life. Fang, *et al.* (1996), for example, explore the high rate of mortality from cardiovascular causes among blacks in New York City, finding that there is substantial variation among blacks based on their place of birth and that, in particular, Southern-born blacks had much higher rates of mortality from cardiovascular disease, and Caribbean-born blacks had much lower such rates, than those of their Northeastern-born counterparts. Barker (1990 and 1995) famously argues that influences on adult health status and mortality extend back to *in utero* nutritional conditions.

One important strand of literature seeks to provide evidence about the nature of black-white mortality gap by following cohorts of black and white individuals, examining the correlation of life-course factors and survival. An extremely important example is Hayward and Gorman's (2004) empirical work using the National Longitudinal Survey of Older Men (NLS-OM). In that paper, the authors show that age-conditioned mortality rates are substantially higher for black men than white men, and the measured early-life social and economic conditions are responsible for only a modest amount of that gap.

An important problem with longitudinal data such as the NLS-OM is a particular form of measurement error—missing data on death. In the NLS-OM data were recorded for men

[1] More broadly, there is a large literature devoted to untangling the relationships between growth in income, improvements in nutrition, increases in education, and improvements in public health and health outcomes (morbidity and mortality). Deaton (2006) gives a valuable assessment of core issues, and Cutler, Deaton, and Lleras-Muney (2006) provide a historical overview. See also Oeppen and Vaupel (2002), Preston (2007), Case, *et al.* (2002). Work on the black-white gap in the U.S. includes Behrman *et al.* (1991) and Elo and Drevenstedt (2006).

aged 45-59 beginning in 1966 and then for several additional years up through 1983. Then in 1990 data were matched with death certificate records for the purpose of improving the precision of the data on death. As we will show below, there is clear evidence that in the construction of the data some deaths were not recorded. In this paper we investigate the problems that arise from this form of measurement error.

The paper proceeds as follows: In Section 2 we investigate theoretical issues of measurement error in a very simple survival model. Section 3 shows that measurement error in death is in fact an important problem in the NLS-OM data, and demonstrates that the way in which this error is handled has a significant impact on inferences one draws from the data about the black-white mortality gap. Section 4 provides a discussion.

2 Measurement Error in Mortality

Our goal in this section is to explore the consequences of measurement error in mortality for estimation of regression-based models of survival. The basic idea of such empirical exercises, using either cross-sectional data or longitudinal data, is simple. Data are collected for a sample of individuals (e.g., age, race, family background, etc.) who, obviously, are alive at the time data are initially collected. Subsequently, deaths are recorded for some individuals in the sample. Then regression analysis is used to examine the statistical correlates of death. Our concern is the mismeasurement of death.

The mismeasurement of deaths of course works in one of two ways—deceased individuals could be classified as being alive *or* deaths could be recorded for those who are alive—but in many data sets most errors might will be one-sided. Consider, for example, the NLS-OM data. All men in these data were interviewed in 1966 (and most were interviewed in subsequent years). Clearly, these individuals were alive at the time data were collected. Data were in 1990 matched with state vital records departments to determine dates of deaths for those who were deceased. It is likely that death certificates were issued only for those who had died. The measurement error in this case is likely mostly one-sided: For some men who died the deaths may not have been matched; some deaths are unrecorded. We focus here on this sort of mismeasurement. Extensions to included both forms of error are straightforward.

2.1 A Survival Model with Age as the Only Covariate

To set the basic idea we start with the simplest possible discrete survival model. We imagine that we observe people with two ages, 0 and 1, and we are interested in the impact of age on the probability of death. A common specification, which we will use here, is that the log of the death rate d_i is linearly dependent on age A_i:

$$d_i = \alpha_0 + \alpha_1 A_i + \varepsilon_i. \tag{1}$$

In estimating equation (1) we would use $ln(D_i/n_i)$ as the dependent variable, where D_i is the number of deaths observed for individuals aged i (assuming D_i is always greater than 0) in a particular discrete time period, and n_i is the number of individuals aged i in the initial survey.

In typical empirical applications one would also include additional covariates. (We turn to that issue shortly.) Also, in typical applications estimation would not proceed with OLS estimation—but instead with some more advanced procedure. Here, though, we work with OLS because by so doing we can most easily highlight the problems that arise if we encounter measurement problems with D_i.

2.1.1 OLS Estimation with No Measurement Error

As a baseline, suppose that in fact deaths are accurately recorded in our data. Then the OLS estimator of our model's key parameter, α_1, is

$$\hat{\alpha_1} = \hat{d_1} - \hat{d_0}, \tag{2}$$

where $\hat{d_0}$ and $\hat{d_1}$ are sample log death rates at ages 0 and 1 respectively.

As is well known, the OLS estimator $\hat{\alpha_1}$ is a consistent estimator for α_1.

2.1.2 OLS Estimation with Measurement Error, Cross-Sectional Data

Our concern here is that some deaths are unrecorded. In particular suppose that proportion q_0 of deaths at age 0 are unrecorded and proportion q_1 of deaths at age 1 are unrecorded. We are interested in the impact our OLS estimator of α_1 for two cases: Cross-sectional data and longitudinal data. We begin here with the case of cross-sectional data.

Here we observe n_0 and n_1 people at the beginning of a period and then observe deaths recorded for the period in each age group. The observed deaths though, are only a subset of actual deaths. In particular, we observe $\tilde{D_0} = (1 - q_0)D_0$ deaths for young individuals and $\tilde{D_1} = (1 - q_1)D_1$ for older individuals. If we simply treat observed deaths as relevant data, it is a matter of simple algebra to verify that our OLS estimator now gives

$$\tilde{\alpha_1} = \hat{\alpha_1} + [ln(1 - q_1) - ln(1 - q_0)], \tag{3}$$

where $\hat{\alpha_1}$ is the consistent estimator from (2).

Comparison of (2) and (3) shows that in general our estimator is not consistent. The direction of the bias depends on the relationship of q_0 to q_1, and does so in an intuitive way. For example, if $q_0 > q_1$, i.e., we miss a higher fraction of deaths among those who die at young ages. Then $ln(1 - q_1)$ will be smaller, in absolute value, than $ln(1 - q_0)$, and our OLS estimator, $\tilde{\alpha_1}$, is biased upward (so we overestimate the impact of aging on death). Of course, if $q_1 \cong q_0$, our estimator will not be too far off.[2]

[2] In this latter case, though, the OLS estimator of α_0—the baseline mortality rate—will be inconsistent. For some applications this parameter may be of less interest.

Our derivations focus here on the case in which some deaths go unrecorded, i.e., the relevant case for our empirical example below. Clearly, though, the logic also applies if opposite pertains; our derivations allow for q_0 and q_1 to be negative (which would occur if some individuals are incorrectly recorded as deceased).

2.1.3 OLS Estimation with Measurement Error, Longitudinal Data

More interestingly, suppose we are working with longitudinal data. In particular, we begin with a sample of n_0 young people. We then observe some deaths in the first period, i.e., observe $\tilde{D_0} = (1 - q_0)D_0$. Then in the next period we observe some additional deaths from these same people when they are one period older, i.e., observe $\tilde{D_1} = (1 - q_1)D_1$. Now our inference problem is slightly more complicated. The reason is that we are mismeasuring not only the number of deaths, but we are also mismeasuring the number of older individuals (i.e., n_1) that we use as the denominator of our death rate among those aged 1.

It is a matter of simple algebra to show that the OLS estimator for this case is now

$$\check{\alpha}_1 = \hat{\alpha_1} + \left[ln(1 - q_1) - ln(1 - q_0) + ln \left(\frac{n_1}{n_1 + q_0 D_0} \right) \right], \tag{4}$$

where again $\hat{\alpha_1}$ is the consistent estimate from (2). Notice that in this case, even if the error rates in reporting deaths are the same for those aged 0 and 1, our estimator is inconsistent. With longitudinal data, when $q_1 \cong q_0$, our estimator is biased downward since the last term in equation (4) is negative. This source of bias will be quite small (in absolute value) if the number of deaths among the young (D_0) is small relative to the number surviving to the older age (n_1), but will be more substantial in a population with a higher death rate among the young.

2.2 A Survival Model with Two Covariates, Age and Race

Suppose now we have two covariates in our model—"age" and "race" in our example. In particular, suppose

$$d_i = \alpha_0 + \alpha_1 A_i + \alpha_2 B_i + \varepsilon_i, \tag{5}$$

where the additional covariate, B_i, is a race indicator,

$$B_i = \begin{cases} 1 & \text{if the respondent is black, and} \\ 0 & \text{if he is white.} \end{cases} \tag{6}$$

Now we let $n_0 = n_0^B + n_0^W$ be the number of individuals aged 0 in our samples, with n_0^B indicating the number of blacks and n_0^W indicating the number of whites. We use analogous notation for those aged 1, $n_1 = n_1^B + n_1^W$.

2.2.1 OLS Estimation with No Measurement Error

With a bit of algebra one can verify that the OLS estimator of the age coefficient α_1 is the weighted sum, for blacks and whites respectively, of the differences between the mean log death rates of the old and the young. Specifically

$$\hat{\alpha_1} = \phi(\hat{d}_1^B - \hat{d}_0^B) + (1-\phi)(\hat{d}_1^W - \hat{d}_0^W), \tag{7}$$

with the weight ϕ given by the surprisingly involved expression,

$$\phi = \frac{(n_0^W + n_1^W)^2 + n_0^B(2n_0^B + n_0^W) + n_1^B(2n_1^B - n_0^W) + n_1^W(n_1^B - n_0^B)}{(n_0^B + n_0^W)^2 + (n_0^B + n_1^B)^2 + (n_0^B - n_1^W)^2 + (n_0^W + n_1^W)^2 + (n_1^B + n_1^W)^2 + (n_0^W - n_1^B)^2}.$$

Similarly, the race coefficient is the weighted sum of the difference between the log death rate of blacks and whites for those aged 0 and the difference between the log death rate of blacks and whites aged 1:

$$\hat{\alpha_2} = \theta(\hat{d}_0^B - \hat{d}_0^W) + (1-\theta)(\hat{d}_1^B - \hat{d}_1^W), \tag{8}$$

with

$$\theta = \frac{(n_1^B + n_1^W)^2 + n_0^B(2n_0^B + n_1^B) + n_0^W(2n_0^W - n_1^B) + n_1^W(n_0^W - n_0^B)}{(n_0^B + n_0^W)^2 + (n_0^B + n_1^B)^2 + (n_0^B - n_1^W)^2 + (n_0^W + n_1^W)^2 + (n_1^B + n_1^W)^2 + (n_0^W - n_1^B)^2}.$$

These estimators are consistent.

As above, our interest is in the impact on the estimators of measurement error in death rates. We let q_0^B be the fraction of deaths that go unreported for blacks aged 0, and define q_0^W, q_1^B, and q_1^W analogously. We generally will be analyzing cases in which these are positive (i.e., deaths are under-reported), though nothing in our derivations is changed if they are negative (i.e., if deaths are over-reported).

2.2.2 OLS Estimation with Measurement Error, Cross-Sectional Data

Suppose again that we have a data set in which we observe initially samples of blacks and whites aged 0 and 1, and then observe deaths in one period. With a bit of algebra we can show that impact of measurement error in deaths is quite intuitive:

$$\tilde{\alpha_1} = \hat{\alpha_1} + \phi[ln(1-q_1^B) - ln(1-q_0^B)] + (1-\phi)[ln(1-q_1^W) - ln(1-q_0^W)], \tag{9}$$

and

$$\tilde{\alpha_2} = \hat{\alpha_2} + \theta[ln(1-q_0^B) - ln(1-q_0^W)] + (1-\theta)[ln(1-q_1^B) - ln(1-q_1^W)], \tag{10}$$

where $\hat{\alpha_1}$ and $\hat{\alpha_2}$ are the consistent estimators given in (7) and (8) respectively.

Equations (9) and (10) provide useful insight into how measurement error is likely to affect our estimates.

Notice, first of all, that if the measurement error is similar for our four age-race groups, $q_0^B \cong q_1^B \cong q_0^W \cong q_1^W$, the OLS estimators will be close to consistent.[3]

Second, our observations about bias to the age coefficient from subsection 2.1.2 carry over here. For instance, if for both blacks and whites deaths are under-reported at higher rates for the young than the old, i.e., if $q_0^B > q_1^B$ and $q_0^W > q_1^W$, then the OLS estimator of α_1 will be biased upward. If, on the other hand, deaths are under-reported at higher rates for the old, the estimator will be biased downward.

Third, it is easy to see how bias might be generated in the race coefficient, $\tilde{\alpha_2}$. Below we show that in the death certificate data matched to the NLS-OM, deaths are under-reported at highest rates for blacks. Suppose, therefore, that $q_0^B > q_0^W$ and $q_1^B > q_1^W$. Clearly, from (10), we can see that the consequence will be that $\tilde{\alpha_2}$ will be biased downward.

2.2.3 OLS Estimation with Measurement Error, Longitudinal Data

We next turn to the more interesting case of longitudinal data. In particular, we suppose that we now are following a cohort of n_0^B black individuals and n_0^W white individuals over two periods, and using those data to estimate our model.

After extensive algebraic manipulation, we can show now the OLS estimator of the age coefficient is

$$\begin{aligned} \breve{\alpha}_1 \;=\; & \left\{ \tilde{\phi}(\hat{d}_1^B - \hat{d}_0^B) + (1-\tilde{\phi})(\hat{d}_1^W - \hat{d}_0^W) \right\} + \\ & \left\{ \tilde{\phi}[ln(1-q_1^B) - ln(1-q_0^B)] + (1-\tilde{\phi})[ln(1-q_1^W) - ln(1-q_0^W)] \right\} + \\ & \left\{ \tilde{\phi}\left[ln\left(\frac{n_1^B}{n_1^B + q_0^B D_0^B} \right) \right] + (1-\tilde{\phi})\left[ln\left(\frac{n_1^W}{n_1^W + q_0^W D_0^W} \right) \right] \right\}, \end{aligned} \tag{11}$$

with weights constructed using

$$\tilde{\phi} = \frac{\begin{array}{c}[(n_0^W + n_1^W)^2 + n_0^B(2n_0^B + n_0^W) + n_1^B(2n_1^B - n_0^W) + n_1^W(n_1^B - n_0^B)] + \\ \{[(4n_1^B - n_0^W + n_1^W)q_0^B D_0^B + (-n_0^B + n_1^B + 2n_0^W + 2n_1^W)q_0^W D_0^W] + [2(q_0^B D_0^B)^2 + (q_0^B D_0^B)(q_0^W D_0^W) + (q_0^W D_0^W)^2]\}\end{array}}{\begin{array}{c}[(n_0^B + n_0^W)^2 + (n_0^B + n_1^B)^2 + (n_0^B - n_1^W)^2 + (n_0^W + n_1^W)^2 + (n_1^B + n_1^W)^2 + (n_0^W - n_1^B)^2] + \\ \{2(n_0^B + 3n_1^B - n_0^W + n_1^W)q_0^B D_0^B + 2(-n_0^B + n_1^B + n_0^W + 3n_1^W)q_0^W D_0^W + [3(q_0^B D_0^B)^2 + 2(q_0^B D_0^B)(q_0^W D_0^W) + 3(q_0^W D_0^W)^2]\}\end{array}}.$$

Three terms in curly brackets on the right hand side of (11) are relatively easy to interpret:

The first term is very similar to the OLS estimator (7). The difference is that the weights have now changed. This expression is thus a consistent, but not efficient, estimator of α_1.

The second term introduce bias in roughly the same way in cross-sectional data, as shown in (9). The only difference is that the weights differ. So if, for example, error rates are similar for old and young individuals in the sample (for both races), $q_0^B \cong q_1^B$ and $q_0^W \cong q_1^W$, then this source of bias will be close to zero.

[3]The estimator of the intercept in our regression $\tilde{\alpha_0}$ will typically be inconsistent, but as we note above, that parameter might be of less interest.

The third term appears for the following reason: If we miss some deaths at age 0, this will cause us to overestimate the base for calculating the death rate at age 1. This in turn causes us to underestimate mortality at age 1; this third term is negative. As in the simpler case above, the source of bias will be quite small (in absolute value) if deaths at age 0 (D_0^B and D_0^W) are infrequent relative to the number of survivors (n_1^B and n_1^W).

As for the coefficient on race, α_2, we can show that the OLS estimator here is

$$\begin{aligned}\breve{\alpha}_2 \;=\; & \left\{\tilde{\theta}(\hat{d}_0^B - \hat{d}_0^W) + (1-\tilde{\theta})(\hat{d}_1^B - \hat{d}_1^W)\right\} + \\ & \left\{\tilde{\theta}[ln(1-q_0^B) - ln(1-q_0^W)] + (1-\tilde{\theta})[ln(1-q_1^B) - ln(1-q_1^W)]\right\} + \\ & \left\{(1-\tilde{\theta})\left[ln\left(\frac{n_1^B}{n_1^B + q_0^B D_0^B}\right) - ln\left(\frac{n_1^W}{n_1^W + q_0^W D_0^W}\right)\right]\right\}, \end{aligned} \tag{12}$$

with weights constructed using

$$\tilde{\theta} = \frac{\begin{matrix}[(n_1^B + n_1^W)^2 + n_0^B(2n_0^B + n_1^B) + n_0^W(2n_0^W - n_1^B) + n_1^W(n_0^W - n_0^B)] + \\ \{[(n_0^B + 2n_1^B - n_0^W + 2n_1^W)q_0^B D_0^B + (-n_0^B + 2n_1^B + n_0^W + 2n_1^W)q_0^W D_0^W)] + (q_0^B D_0^B + q_0^W D_0^W)^2\}\end{matrix}}{\begin{matrix}[(n_0^B + n_0^W)^2 + (n_0^B + n_1^B)^2 + (n_0^B - n_1^W)^2 + (n_0^W + n_1^W)^2 + (n_1^B + n_1^W)^2 + (n_0^W - n_1^B)^2] + \\ \{2(n_0^B + 3n_1^B - n_0^W + n_1^W)q_0^B D_0^B + 2(-n_0^B + n_1^B + n_0^W + 3n_1^W)q_0^W D_0^W + [3(q_0^B D_0^B)^2 + 2(q_0^B D_0^B)(q_0^W D_0^W) + 3(q_0^W D_0^W)^2]\}\end{matrix}}.$$

Again we have three terms in curly brackets to interpret:

The first term is similar to the OLS estimator (8), but with different weights. This term is a consistent, but inefficient, estimator of α_2.

The second term is similar to bias identified in the cross-sectional case, as shown in (10). As we mention above, in the empirical example that we pursue mismeasurement is a bigger problem for blacks than for whites, i.e., $q_0^B > q_0^W$ and $q_1^B > q_1^W$. For such a situation this second term is clearly negative; this biases our OLS estimator downward.

The third term has an ambiguous sign. Both expressions within the square bracket are negative. We can see what happens in some special cases. For instance, if there is measurement error for blacks but not whites, $q_0^B > 0$ and $q_0^W = 0$, the entire term is negative. We thus expect more generally that the entire term is negative as long as the measurement error for blacks is sufficiently larger than the measurement error for whites. Also, we notice that the term is more likely to be negative when D_0^B is large relative to D_0^W. Importantly, since the weights in our expression depend on $\tilde{\theta}$, the composition of the population (e.g., the proportion black) affects the size of this bias. As above, the size of this bias will be small (in absolute value) if the number of deaths (D_0^B and D_0^W) are infrequent relative to the number of survivors (n_1^B and n_1^W).

3 An Empirical Example Using the NLS-OM

3.1 The National Longitudinal Survey of Older Men

The National Longitudinal Surveys (NLS) are a set of surveys designed to gather information at various points in time on labor market activities and other significant life events of several groups of people. One of these surveys, the National Longitudinal Survey of Older Men (NLS-OM) has proved to be an important source of research in social science, including the analysis of the role of race and socioeconomic status on mortality. Indeed, a number of recent papers use these data for that purpose, including Hayward, *et al.* (1997), Hayward and Gorman (2004), and Warner and Hayward (2006).

The NLS-OM was first administered to 5020 respondents by interviewers from the United States Census Bureau in 1966. The age eligibility was men 45 to 59 on April 1, 1966. Thus, the survey covers birth cohorts from 1906 through 1921. The survey was repeated a further 12 times between the years of 1967 and 1983. An additional version of the survey was completed by living respondents or by the widows or other family members of deceased respondents in 1990. At that time, the oldest cohorts in the NLS-OM were aged 80 or older, and there was therefore substantial mortality; mortality rate could be inferred up through fairly old ages for these men. As of the last date of data collection, it appears that 53.3% of the older men samples were deceased.

Importantly, for the work that follows, the NLS older men survey obtained mortality information for the 1966-1990 periods on 2674 deaths in two ways. First, the 1990 data reports mortality via death certificates from state vital records departments. Second, throughout the data collection process—up through 1983 and again in 1990—there are reports by widows or next-of-kin that can be used to infer death.

Before turning to an analysis of mortality and the correlates of mortality, we report basic statistics about the socio-economic characteristics of the men in the sample. The childhood conditions of NLS older men reflected the characteristics of the American population during the early twentieth century. Tables 1 and 2 show that these men generally lived in households where the heads had low levels of education by current standards. Most of them lived with both their biological mother and father, though some did live in homes in which one or both parents were absent. For the most part, NLS older men respondents grew up in rural communities, many lived in towns with populations of fewer than 25,000 people or lived in rural farm areas. Although these men were better educated than their parents, the amount of schooling they completed was low compared with present averages. As adults, these men were much more likely to live in or near urban areas than as children. The data included some lifestyle measures, including alcohol consumption (which is quite low among those who provide reports), smoking, and the body mass index (BMI), which can be used to assess the subsequent impact on mortality of obesity.

3.2 Initial Regression Analysis Using Vital Record, 1966–1983

Our primary interest here is to draw statistical inference about the role of race as a correlate of mortality in the sample of NLS older men. As we mention above, such analysis has already been undertaken, most notably in the important work of Hayward and Gorman (2004). We revisit this analysis because we hope to understand the role of measurement error in mortality as discussed above.

As we have noted, in the NLS-OM data, death can be recorded in one of two ways: By death certificates from state vital records or by an indication that data went unrecorded because the respondent was deceased.

We begin here by undertaking regression analysis of mortality using the entire sample over the 1966-1983 period, and taking the "conservative" approach of treating each respondent as alive unless a death is recorded by death certificate. Notice that in taking this approach we are replicating the analysis that a research would undertake who had an initial sample of individuals who were known to be alive at a point in time (in our case 1966) and then had access to official records that recorded deaths for that sample. Notice that to the extent that there are deaths for which death certificates are *not* successfully matched to the original data, deaths will be under-reported using this approach.

For our first set of analyses, we restrict attention to the period 1966 through 1983. The reason for doing focusing on this period is that over this span, regular data collection continued for the NLS-OM cohorts. Thus for this period we can draw some reasonable inferences about the consequences of taking the "conservative" approach of using vital statistics as a means of recording deaths.

The basic regression approach we take follows Warner and Hayward's (2006) paper, "Early-Life Origins of the Race Gap in Men's Mortality." Thus we take a digression here to set out the framework. In that paper, the authors conduct survival analysis (also called "event-history analysis") in which a series of discrete-time hazard models are estimated for the purpose of evaluating the ways in which social and economic conditions in childhood are associated with mortality. The specific goal is to see how those early life conditions contribute to the race gap in men's mortality.[4] Their analysis is conducted by estimating a series of models that regress the risk of mortality on each of several sets of early-life condition separately. Through the changes in the coefficients across models, one can potentially assess the life-course pathways that account for the race gap in mortality. The authors argue that early life conditions indirectly affect the race gap in mortality via adult socioeconomic status.

For the regression models $h(a) = \lim_{n \to 0} \frac{P(a+n>T\geq a|T\geq a)}{n}$ gives the force of mortality at exact age a, given that a person has survived to that age. The basic association between mortality risk and age is then assumed to follow the log-linear model that we employed above

$$lnh(a) = \beta_0 + \beta_1 A, \tag{13}$$

[4]See also Allison (1984), Hayward and Gorman (2004), and Castilla (2007).

where A is the age of a person at his previous birthday. The series of nested models are Model 1:

$$lnh(a) = \beta_0 + \beta_1 A + \gamma_i CHILD_i,$$

where $CHILD$ represents a key set of characteristics established in childhood—race, five-year birth cohort, and being foreign-born—that are included in every regression. Then Model 2 is

$$lnh(a) = \beta_0 + \beta_1 A + \gamma_i CHILD_i + \beta_2 EDUC_i,$$

where $EDUC$ represents education of head of household when that respondent was a child; Model 3 is

$$lnh(a) = \beta_0 + \beta_1 A + \gamma_i CHILD_i + \beta_2 FAMILY_i,$$

where $FAMILY$ is a vector that represents family structure; and Model 4 is

$$lnh(a) = \beta_0 + \beta_1 A + \gamma_i CHILD_i + \beta_2 COMM_i,$$

where $COMM$ represents community characteristics.

The first model is the "baseline model" which gives the main effect of age and race on mortality. Then the intention of the analysis is assess whether parental education ($EDUC$), family structure ($FAMILY$), and community characteristics ($COMM$) in childhood affect the magnitude of the key parameter estimates.

We use the full sample of 5020 for this analysis, excluding only a small number of cases for which we lack data on the independent variable.[5] The results of the analyses are given, in full, in the appendix.[6] Our interest here is primarily on the coefficients on age and race, so we present a truncated version of the results in Panel A of Table 3. As expected, the log force of mortality is increasing in age. Also, as expected, conditional on age (and cohort of birth and foreign-born status) the force of mortality is substantially higher for blacks than for whites. We also find that the estimated effect of race on mortality is not substantially altered by inclusion of other measured individual-level characteristics.

3.3 Regression Analysis Using Deaths Reported in the NLS, 1966–1983

As we have noted, use of vital records to record deaths is a very conservative approach here, and is likely to lead to under-reporting of deaths. As we have seen above, this sort of mismeasurement can create serious bias in parameter estimates, especially if the unrecorded deaths are systematically related to characteristics of individuals in the sample.

[5]Thus, for example, we are missing data on age, cohort, foreign-born status and/or race on 48 individuals, which gives us a final sample size of $n = 4972$ for Model 1.

[6]For the results, we report the estimated coefficients instead of the hazard ratios.

To see how this might matter here, we repeat our regression analysis, but now we include not only deaths that are recorded in vital records, but also those that appear in the NLS data collection process. Thus, for example, for many records, data are missing for a particular year (and all subsequent years) and the recorded reason is that the respondent is deceased. For a large number of such cases, the respondent also has a death recorded via death certificate. For some cases, though, these deaths were not recorded in the vital statistics. We ask what happens if these cases are included.

Results are reported in Panel B of Table 3. The differences between Panel A and Panel B are striking. We notice that if we use the death certificates to record deaths, we substantially *over-estimate* the age coefficient and substantially *under-estimate* the race coefficient. The theoretical results from the previous section give clues about how this problem might arise.

The data used to estimate our key regressions are of course "panel data"; identification of the parameters comes in part from the longitudinal component (as measured mortality changes as cohorts age) and from the cross-section (as measured mortality varies in the cross-section for men of different ages). Thus the lessons from the longitudinal analysis and the cross-sectional analysis pertain. Fortunately, those lessons are very similar, as can be seen from comparing results in (9) and (10) with results in (11) and (12). In particular, the nature of the bias of the key coefficients is related to the extent to which the individual characteristics of people with omitted deaths differ from those with recorded deaths.[7]

3.4 Characteristics of Individuals with Likely Deaths *Not* Recorded in Vital Records

To repeat our key point from the analysis in the previous section, the omission of death records is likely to be particularly problematic if the characteristics to those whose deaths are unrecorded differ from the characteristics of those whose deaths are recorded. With this

[7]This regression includes as "deaths" those deaths that are recorded by death certificates *and* deaths that appear also as recorded deaths in the NLS-OM data. It is worth noting that for a few of these cases the age of death does not line up with the reports. For example, 133 individuals whose death certificates show that they were dead before 1983 with no death records provided in NLS-OM are treated as dead before 1983 in the regression. For the 80 individuals whose death certificates and NLS-OM records confirm that they were alive before 1983, but are recorded as dead after the year of their last interview (1990) according to death certificates, we assume they were alive until 1990. The 7 respondents whom are twice confirmed of being dead before 1983 from NLS-OM and death certificates but are also recorded as dead before the year of first interview (1966) from death certificates are treated as died in 1967 in this regression. The 54 respondents whom are confirmed of being dead before 1983 by the NLS-OM death records but reported dead between 1983 and 1990 from death certificates are treated as alive before 1983 and dead between 1983 and 1990. Also, within this group, there is one case reported dead after 1990 according to death certificates; he is treated as dead in 1990. We suspect that most of the deaths that show up in NLS-OM data are in fact deaths, as it is difficult to see why widows or other surviving relatives would mis-report this. In any event, while it is possible that some of these cases are reported deaths for individuals who are in fact alive, we think we are much closer to actual outcomes by including these cases.

in mind, consider Table 4. This table gives average characteristics for two groups—those who had deaths recorded in the vital records and matched to the NLS-OM data, and those who had deaths as indicated in the NLS *but not in the death certificate data from state vital records.* This latter group of 382 men are those that would be omitted from an analysis based on death certificate data.

We notice some striking results in comparing the two samples. First, the death ages recorded in NLS-OM (collected directly from the widows or proxy of the respondents) indicate that the group of 382 men died at young ages. Their average death age is 62.02, while it is 62.95 for the other group. Among blacks, the 382 men seem to have died at younger ages. The average death age is about 61.55 for the black within the 382 men, while the average death age is about 62.88 for the blacks in the other group. Second, this group of 382 men is disproportionately black. It is found that the proportion of blacks in the 382 men is larger than the proportion of blacks in the other group. There are 16.22% blacks in the 382 men, while there are only 10.54% blacks in the other group. The above comparisons between the two samples raise our concern that the unrecorded deaths differ systematically across groups. The measurement error is highly systematic. Thus, the empirical evaluation of black-white mortality gap should account for the presence of measurement errors. Since we probably lose more deaths for those more likely to be black and young at death age, the age coefficients are over-estimated and the race coefficients are under-estimated if the omitted deaths are not taken into consideration. The differences in the age and race estimates between Panels A and B in Table 3 are mostly likely due to the systematically unreported deaths.

3.5 Regression Results Using Consistent Records Only

The omission of deaths that go unreported in the death certificate data from regression analyses give us results that differ substantially from those that include such deaths. As we have noted, though, some problems remain. We cannot know for certain, for example, that deaths that have no death certificates did actually occur. Also, there are cases in which the death ages do not line up from the two data sources.

One way of dealing with these inconsistencies is to use only data for which we have consistent records. In empirical work generally researchers often discard data for which records are incomplete or inconsistent. Here we can follow in that tradition by doing the same for records for which death records and ages fail to match up consistently.[8]

Results of this exercise are reported in Panel C of Table 3. The most important point to make with this set of results is that the key estimated coefficients are in between those found in Panel A and those found in Panel B.

Our interpretation is as follows: As noted above, the Panel A estimates are deeply flawed because a substantial number of deaths are going unreported, and moreover because the

[8]By following this path, we are, we believe, roughly following Hayward and Gorman (2004) and Warner and Hayward (2006).

characteristics of men with unreported deaths differ substantially from those with reported deaths. In particular, we have seen that those with unreported deaths are more likely to be black and young at age of death. We suspect that the results in Panel B are likely to be more accurate.

Now the sample used to produce the results reported in Panel C excludes a number of cases that seem "problematic," but in so doing once again excludes a large number of deaths that almost certainly occurred, and as we have just emphasized the excluded cases are disproportionately deaths of those who are black and young at age of death. At least in the data used for Panel C we are not treating those cases as individuals who survived throughout the period. Still, as an empirical matter, simply excluding those cases is almost as bad as coding them as survivors. It appears that we end up with badly inconsistent estimates of our key parameters.

3.6 Regression Results Using Data 1966-1990

In the analyses we have just reported, we used data only up through 1983. By so doing we can identify a number of likely deaths that are *not* recorded in the death certificate data, because regular interviews on the men were being conducted. This allows us further to see how inferences differ if we (properly) included those as deaths or (improperly) simply exclude those cases from analysis.

Previous research using these data up through 1990, which is clearly advantageous because this allows for the inclusion of deaths at older ages. Of course, the basic measurement issues still pertain for these years, even though we do not have regularly-collected data with which to examine the problem. What we do have, though, is one additional final report in 1990 in which data were again collected from men who were alive or from widows or other relatives for those who were deceased.

Thus we can repeat our regression analyses, again treating death records in three different ways. First, we can rely on death certificate records only. Second, we can included deaths reported on death certificates *and* other deaths reported in the NLS-OM. Third, we can simply exclude all cases with inconsistent records. Given our discussion above, we suspect the second of these options is likely to produce correct inferences.

With this in mind, consider results reported in Table 5. The basic pattern is similar to that reported in Table 3. Most importantly, compare results in Panels B and C. If we exclude inconsistent records, we likely under-estimate the impact of race on mortality and over-estimate (by a small amount) the impact of age on the force of mortality.

When comparing our results from Table 5 to the results from Warner and Hayward's (2006) paper, it seems that results reported in Panel C are much closer to their results. The estimates for age and race in Panel C are most comparable to their estimates for age and race. Results in Panel C reveal a similar story for the effects of childhood characteristics on men's mortality as in their paper. We think that the sample excluding all cases with inconsistent records is likely to be the most comparable one to the sample in their paper, although the impact of race would be under-estimated and the impact of age would be over-estimated on mortality when using this sample. The slight disagreements between our results and results in their paper imply that we are not using the same samples since their scheme used to deal with the unrecorded deaths is still unknown.

4 Conclusion

A large and growing literature seeks to understand the role of life-course events, stretching back through early childhood. The analysis of such issues typically requires the use of longitudinal data. Such data is subject to a variety of problems, one of which we discuss here: Often we will have incomplete records of deaths for a sample.

The initial contribution of the paper is to study the nature of the biases that are introduced when researchers face measurement error in mortality. Using a very simple model, we are able to make some useful observations. For example, we see:

First, if the source of identification in the model is cross-sectional, reasonably consistent inference of key parameters in the regression might be possible as long as the mismeasurement of mortality is the same for all key demographic groups (e.g., if unrecorded deaths are not related to age or race).

Second, in data in which identification comes in part from longitudinal variation, the "age coefficient" in a survival regression is likely to be biased downward even when the age-specific rate of measurement error is the same across ages.

Third, if deaths are under-recorded at higher rates for one racial group than another, we are likely to under-estimate the role of race on mortality outcomes for the under-recorded group.

With these lessons in mind, we also provided an empirical application of mismeasurement in deaths. Our example comes from the NLS older men data (or death certificate data). We found that in the 1990 data, which are widely used for analysis, there are a fairly large number of men who likely died for whom there was no matched death certificate data. These "omitted deaths" in the data are clearly non-random. In general, blacks are much more likely to be in this group than whites. This is especially true of black men who die at young ages. We suspect that NLS responses contain less measurement error than death certificates. Inclusion of these deaths in the analysis proves to move estimated parameters by a fair amount.

One very important point in our work is that researchers cannot hope to get rid of biases introduced by mismeasurement by adopting the seemingly sensible rule of "excluding inconsistent records." When we restrict attention only to data that have completed records we often will be excluding cases that are *not* missing at random, and biases can thereby introduced.

At this point, it seems likely that the only way to make additional headway is to work to improve the quality of the data. Fortunately, it is likely that that will be possible. At this point, even the youngest of the men surveyed in the NLS-OM data (those aged 45 in 1966) are now quite old, approximately 90 years old, and those who were older would be as old as 105. In short, most of these men are now deceased. Thus by matching the data once again to vital records it will be possible to determine the age of death for almost every one for whom death was recorded in vital records, and then see who simply was missed in the records. This leads to a different (and interesting) inference problem, but one that is much easier to handle than the one encountered with the data as they are now constructed.[9]

[9]It is possible that we will confirm that many of the missing cases arise because of incorrectly recorded Social Security Numbers (SSNs). One possibility is that many of the men did not get SSNs until they were older or perhaps were never issued a SSN. Again this can be checked.

5 References

Allison, Paul, 1984. *Event History Analysis: Regression for Longitudinal Event Data.* California: Sage.

Almond, Douglas, 2006. "Is the 1918 Influenza Pandemic Over? Long-Term Effects of In Utero Influenza Exposure in the Post-1940 U.S. Population," *Journal of Political Economy,* 114(4), 672-712.

Barker, D. J. P., 1990. "The Fetal and Infant Origins of Adult Disease," *British Medical Journal,* 301, 1111.

Barker, D. J. P., 1995. "Fetal Origins of Coronary Heart Disease," *British Medical Journal,* 311, 171-74.

Behrman, Jere R., Robin Sickles, Paul Taubman, and Abdo Yazbeck, 1991. "Black-White Mortality Inequalities," *Journal of Econometrics*, 50(1-2), 183-203.

Bengtsson, T. and M. Lindstrm, 2000. "Childhood Misery and Disease in Later Life: The Effects on Mortality in Old Age by Hazards Experienced in Early Life, Southern Sweden, 1760-1894," *Population Studies,* 54, 263-77.

Black, Dan, Seth Sanders, and Lowell Taylor, 2003. "Measurement of Higher Education in the Census and Current Population Survey," *Journal of the American Statistical Association,* 98(463), 545-554.

Bollinger, Christopher, 1996. "Bounding mean regressions when a binary regressor is mismeasured," *Journal of Econometrics,* 73, 387-399.

Bound, John, Charles Brown, Greg J. Duncan, and Willard L Rodgers, 1990. "Measurement Error In Cross-Sectional and Longitudinal Labor Market Surveys: Results from Two Validation Studies," *Panel Data and Labor Market Studies,* 1-19, Elsevier Science Publishers B.V. (North-Holland).

Bound, John, Charles Brown, Greg J. Duncan, and Willard L Rodgers, 1994. "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data," *Journal of Labor Economics,* 12(3), 345-68.

Bound, John, Charles Brown, and Mathiowetz, Nancy, 2001. "Measurement Error in Survey Data," in *Handbook of Econometrics* (vol. 5), eds. Heckman, J.J. and Leamer, E.E., Elsevier.

Bound, John and Alan B. Krueger, 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make A Right?," *Journal of Labor Economics*, 9(1), 1-24.

Case, Anne, Darren Lubotsky, Christina Paxson Source, 2002. "Economic Status and Health in Childhood: The Origins of the Gradient," *American Economic Review,* 92(5), 1308-1334.

Castilla, Emilio J., 2007. *Dynamic Analysis in the Social Sciences.* Academic Press and Elsevier.

Costa, Dora L., 2004. "Race and Pregnancy Outcomes in the Twentieth Century: A Long-Term Comparison," *Journal of Economic History,* 64(4), 1056-86.

Costa, Dora L., Lorens Helmchen, and Sven Wilson, 2007. "Race, Infection, and Arteriosclerosis in the Past," *Proceedings of the National Academy of Science,* 104(33), 13219-24.

Cutler, David, Angus Deaton, and Adriana Lleras-Muney, 2006. "The Determinants of Mortality," *Journal of Economic Perspectives,* 20(3), 97-102.

Deaton, Angus, 2006. "The Great Escape: A Review Essay on Robert W. Fogel's 'The Escape from Hunger and Premature Death, 1700-2100'," *Journal of Economic Literature,* 44(1), 106-114.

Elo, Irma T. and Greg L. Drevenstedt, 2006. "Black-White Differentials in Cause-Specific Mortality in the United States during the 1980s: The Role of Medical Care and Health Behaviors," PSC Working Paper Series PSC 06-02.

Elo, Irma T. and Samuel H. Preston, 1992. "Effects of Early-life Conditions on Adult Mortality: A Review," *Population Index,* 58(2), 186-212.

Elo, Irma T. and Samuel H. Preston, 1994. "Estimating African-American Mortality from Inaccurate Data," *Demography,* 31(3), 427-58.

Fang, Jing, Shantha Madhavan and Michael Alderman, 1996. "The Association Between Birthplace and Mortality from Cariovascular Causes among Black and White Residents of New York City," *New England Journal of Medicine,* 335(21), 1545-51.

Fogel, Robert, 1993. "New Sources and new Techniques for the Study of Secular Trends in Nutritional Status, Health, Mortality, and the Process of Aging," *Historical Methods,* 26(1), 5-43.

Fogel, Robert, 2004. *The Escape from Hunger and Premature Death, 1700-2100.* Cambridge University Press.

Fuller, Wayne A., 1987. *Measurement Error Models.* New York: Wiley.

Guralnik, Jack, Kenneth Land, Dan Blazer, Gerda Fillenbaum and Laurence Branch, 1993. "Educational Status and Active Life Expectancy among Older Blacks and Whites," *New England Journal of Medicine,* 239(22), 110-16.

Harper, Sam, John Lynch, Scott Burris and George Smith, 2007. "Trends in the Black-White Life Expectancy Gap in the United States, 1983-2003," *Journal of the American Medical Association,* 297, 1224-32.

Hayward, Mark D. and Melonie Heron, 1999. "Racial Inequality in Active life among Adult Americans," *Demography,* 36(1), 77-91.

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

Hayward, Mark D. , William R. Grady, Melissa A. Hardy, and David Sommers, 1989. "Occupational Influences on Retirement, Disability, and Death," *Demography,* 26(3), 393-409.

Hayward, Mark D., Amy M. Pienta, Diane K. McLaughlin 1997. "Inequality in Men's Mortality: The Socioeconomic Status Gradient and Geographic Context," *Journal of*

Health and Social Behavior, 38(4), 313-330.

Hosmer, David W., Stanley Lemeshow, and Susamme May, 2008. *Applied Survival Analysis: Regression Modeling of Time to Event Data.* New York: Wiley.

Levine, Robert, James Foster, Robert Fullilove, Mindy Fullilove, Nathaniel Briggs, Pamela Hull, Baqar Husaini, Charles Hennekens, 2001. "Life Expectancy, 1933-1999: Implications for Healthy People 2010," *Public Health Reports,* 116, 474-83.

Oeppen, Jim and Vaupel, James W, 2002."Broken Limits to Life Expectancy," *Science*, 296(5570), 1029-1031.

Preston, Samuel H., 2007. "The Changing Relation between Mortality and Level of Economic Development," *International Journal of Epidemiology*, 36(3), 484-490.

Preston, Samuel H., Mark Hill, and Greg Drevenstedt, 1998. "Childhood Conditions that Predict Survival to Advanced Ages Among African Americans," *Social Science Medicine,* 47(9), 1231-46.

Sen, Amartya, 2001. "Mortality as an Indicator of Economic Success and Failure," *Economic Journal,* 108(446), 1-25.

Sorlie, P., E. Rogot, R. Anderson, N. J. Johnson and E. Backlund, 1992. "Black-White Mortality Differences by Family Income," *Lancet,* 340(8815), 346-50.

Vaupel James W., 1997. "The Remarkable Improvements in Survival at Older Ages," *Philosophical Transactions of the Royal Society of London–Series B: Biological Sciences,* 352(1363), 1799-804.

Warner, David F., and Mark D. Hayward, 2006. "Early-Life Origins of the Race Gap in Men's Mortality," *Journal of Health and Social Behavior,* 47, 209-26.

6 Tables

Table 1: Childhood Circumstances of NLS Men

Variable	Percentage
Household Head's Education When the Respondent Was 15	
6 years or less	24.99
7-8 years	20.31
9-12 years	10.77
13 or more years	6.41
Missing	37.52
Household Head's Occupation When the Respondent Was 15	
Professional or military	4.05
Managerial	11.89
Clerical	1.92
Sales	3.06
Crafts	13.47
Operative	12.4
Private household or service worker	5.36
Farmer	32.22
Farm laborer	1.46
Laborer	6.30
Missing	7.88
Foreign Born	6.17
Parent's Nativity	
One parent was foreign born	6.57
Two parents were foreign born	20.95
Neither parent was foreign born	70.54
NA	1.94
Living Arrangement When the Respondent Was 15	
Father and mother	75.17
Father and stepmother	1.73
Mother and stepfather	2.11
Father only	3.13
Mother only	8.68
Other	8.69
NA	0.50
Mother's Work Status When the Respondent Was 15	
Did not work	59.60
Worked	10.88
Missing	29.52
Childhood Urban/Rural Residence	
City with 100,000 or more people	19.79
City with 25,000-100,000 or more people	10.76
Suburb of a large city	2.38
Town with fewer than 25,000 people	27.29
Rural nonfarm area	3.70
Rural farm area	35.54
NA	0.53

[1] Using weighted percentages

Table 2: Characteristics of NLS Men as Adults

Variable	Mean or Percentage
Demographic Characteristics	
Age (mean)	51.55
Black	8.69
Birth Cohort	
1906-1910	27.72
1911-1915	33.34
1916-1921	38.94
Education	
8 years or less	35.42
9-12 years	45.97
13 or more years	18.60
Marital Status	
Married	89.28
Never married	4.57
Divorced	4.23
Widowed	1.92
Urban/Rural Residence	
Urban	49.73
Outside urban	16.52
Rural	33.75
Net Asset (mean)	21717.17
Total Family Income (mean)	7462.867
Body Mass Index	
Under 20	3.29
20-23	13.63
23.1-25	18.01
25.1-27.5	25.45
27.6-52.1	18.02
Missing	21.60
Mean Weekly Alcohol Consumption	
1-2 drinks	19.51
3-4 drinks	5.85
5 or more drinks	5.37
Missing	69.27
Smoking Behavior 1	
Currently smoking	13.09
Currently not smoking	86.91
Smoking Behavior 2	
Never smoked	33.87
Ever Smoked	66.13

[1] Using weighted means and percentages
[2] Net asset and total family income are in dollars; all negative values of net asset and total family income are adjusted to zero

Table 3: Survival Regression Results for Data up to 1983

Panel A. Deaths with Death Certificate Records

	Model 1	Model 2	Model 3	Model 4
Age	0.069**	0.068**	0.069**	0.069**
	(0.018)	(0.018)	(0.018)	(0.018)
Black	0.166**	0.155*	0.159*	0.187**
	(0.063)	(0.065)	(0.065)	(0.064)
Birth Cohort	×	×	×	×
R foreign-born	×	×	×	×
Education of head of household		×		
Family structure			×	
Community size				×
N	4972	4972	4962	4964

Panel B. Including Deaths from NLS Records

	Model 1	Model 2	Model 3	Model 4
Age	0.058**	0.056**	0.058**	0.057**
	(0.016)	(0.016)	(0.016)	(0.016)
Black	0.351**	0.331**	0.339**	0.376**
	(0.054)	(0.055)	(0.055)	(0.055)
Birth Cohort	×	×	×	×
R foreign-born	×	×	×	×
Education of head of household		×		
Family structure			×	
Community size				×
N	4972	4972	4962	4964

Panel C. Discarding Inconsistent Records

	Model 1	Model 2	Model 3	Model 4
Age	0.058**	0.058**	0.060**	0.059**
	(0.020)	(0.020)	(0.020)	(0.020)
Black	0.279**	0.258**	0.265**	0.299**
	(0.067)	(0.068)	(0.069)	(0.068)
Birth Cohort	×	×	×	×
R foreign-born	×	×	×	×
Education of head of household		×		
Family structure			×	
Community size				×
N	4326	4326	4317	4321

[1] Standard errors in parentheses
[2] * significant at 5 %; ** significant at 1%
[3] × corresponds to unreported/statistically insignificant estimates

Table 4: Comparisons between Characteristics of Individuals with Deaths *without* Death Certificates (NLS 382 Men) and Individuals with Deaths *with* Death Certificates (NLS 1116 Men)

	Mean or Percentage	
Variable	382 Men	1116 Men
Demographic Characteristics		
Age (mean)	52.44	52.82
Black	16.22	10.54
Birth Cohort		
1906-1910	35.46	39.07
1911-1915	37.74	34.65
1916-1921	26.80	26.28
Age at Death (mean)	62.02	62.95
Cohort$(1906-1910)$	65.74	67.23
Cohort$(1911-1915)$	62.19	62.11
Cohort$(1916-1921)$	56.50	57.79
Black	61.55	62.88
White	62.11	62.96
Foreign Born	62.40	62.58
Not Foreign Born	61.96	62.99
Education		
8 years or less	42.41	44.39
9-12 years	41.25	43.12
13 or more years	16.34	12.50
Marital Status		
Married	78.25	88.10
Never married	8.89	4.06
Divorced	7.80	5.25
Widowed	5.06	2.59
Urban/Rural Residence		
Urban	55.18	47.43
Outside urban	13.58	17.80
Rural	31.25	34.77
Net Asset (mean)	15320.48	16517.5
Excluding missing values	18914.47	20289.49
Total Family Income (mean)	6233.992	6325.498
Excluding missing values	7653.39	7838.9

[1] Using weighted means and percentages
[2] Net asset and total family income are in dollars; all negative values of net asset and total family income are adjusted to zero
[3] Excluding missing values considers all cases (including all positive and negative values) for net asset and total family income except cases with missing data

Table 5: Survival Regression Results for Data up to 1990

Panel A. Deaths with Death Certificate Records

	Model 1	Model 2	Model 3	Model 4
Age	0.070**	0.070**	0.071**	0.070**
	(0.014)	(0.014)	(0.014)	(0.014)
Black	0.119*	0.107*	0.113*	0.135**
	(0.049)	(0.050)	(0.051)	(0.050)
Birth Cohort	×	×	×	×
R foreign-born	×	×	×	×
Education of head of household		×		
Family structure			×	
Community size				×
N	4972	4972	4962	4964

Panel B. Including Deaths from NLS Records

	Model 1	Model 2	Model 3	Model 4
Age	0.065**	0.063**	0.066**	0.064**
	(0.013)	(0.013)	(0.013)	(0.013)
Black	0.281**	0.262**	0.271**	0.302**
	(0.044)	(0.046)	(0.046)	(0.045)
Birth Cohort	×	×	×	×
R foreign-born	×	×	×	×
Education of head of household		×		
Family structure			×	
Community size				×
N	4972	4972	4962	4964

Panel C. Discarding Inconsistent Records

	Model 1	Model 2	Model 3	Model 4
Age	0.072**	0.071**	0.073**	0.072**
	(0.015)	(0.015)	(0.015)	(0.015)
Black	0.228**	0.208**	0.214**	0.245**
	(0.051)	(0.053)	(0.053)	(0.052)
Birth Cohort	×	×	×	×
R foreign-born	×	×	×	×
Education of head of household		×		
Family structure			×	
Community size				×
N	4326	4326	4317	4321

[1] Standard errors in parentheses

[2] * significant at 5 %; ** significant at 1%

[3] × corresponds to unreported/statistically insignificant estimates

7 Appendix Tables

Table 6: Effect of Early Life Characteristics on Men's Mortality, National Longitudinal Survey of Older Men (1966 – 1990)

	Sample A (up to 1983)			
	Model 1	Model 2	Model 3	Model 4
Age	0.069**	0.068**	0.069**	0.069**
	(0.018)	(0.018)	(0.018)	(0.018)
Black	0.166**	0.155*	0.159*	0.187**
	(0.063)	(0.065)	(0.065)	(0.064)
Cohort (1906 – 1910 reference)				
1911 – 1915	-0.007	-0.008	-0.009	-0.01
	(0.110)	(0.110)	(0.110)	(0.110)
1916 – 1921	-0.1	-0.101	-0.096	-0.106
	(0.192)	(0.192)	(0.192)	(0.193)
R foreign-born	-0.456**	-0.455**	-0.448**	-0.471**
	(0.157)	(0.157)	(0.157)	(0.157)
Education of head of household ($\leq$ 6 years reference)				
Missing		0.039		
		(0.072)		
7 – 8 years		-0.033		
		(0.093)		
9 – 12 years		0.08		
		(0.110)		
13– years		-0.114		
		(0.152)		
Family structure (mother and father reference)				
Father and stepmother			-0.024	
			(0.227)	
Mother and stepfather			0.286	
			(0.180)	
Father only			0.238	
			(0.147)	
Mother only			-0.083	
			(0.103)	
Male relative			0.283*	
			(0.136)	
Other arrangement			-0.008	
			(0.128)	
Living on his own			0.171	
			(0.233)	
Community size (large city 100,000+ reference)				
Smaller city (25,000 – 100,000)				-0.107
				(0.119)
Suburb of a large city				-0.097
				(0.234)
Town less than 25,000				0.153
				(0.089)
Rural, non-farm				0.247
				(0.147)
Rural, farm				-0.039
				(0.085)
Constant	-7.748**	-7.732**	-7.778**	-7.764**
	(1.045)	(1.047)	(1.046)	(1.050)
N	4972	4972	4962	4964

[1] Standard errors in parentheses
[2] * significant at 5 %; ** significant at 1%

Table 7: Effect of Early Life Characteristics on Men's Mortality, National Longitudinal Survey of Older Men (1966 – 1990)

	Sample B (up to 1983)			
	Model 1	Model 2	Model 3	Model 4
Age	0.058**	0.056**	0.058**	0.057**
	(0.016)	(0.016)	(0.016)	(0.016)
Black	0.351**	0.331**	0.339**	0.376**
	(0.054)	(0.055)	(0.055)	(0.055)
Birth cohort (1906 – 1910 reference)				
1911 – 1915	-0.060	-0.065	-0.059	-0.063
	(0.096)	(0.096)	(0.096)	(0.096)
1916 – 1921	-0.203	-0.209	-0.195	-0.214
	(0.167)	(0.167)	(0.167)	(0.167)
R foreign-born	-0.016	-0.014	-0.008	-0.039
	(0.116)	(0.116)	(0.116)	(0.117)
Education of head of household ($\leq$ 6 years reference)				
Missing		0.085		
		(0.062)		
7 – 8 years		-0.042		
		(0.082)		
9 – 12 years		0.050		
		(0.097)		
13– years		-0.093		
		(0.133)		
Family structure (mother and father reference)				
Father and stepmother			-0.129	
			(0.207)	
Mother and stepfather			0.247	
			(0.163)	
Father only			0.109	
			(0.135)	
Mother only			0.009	
			(0.086)	
Male relative			0.180	
			(0.122)	
Other arrangement			0.040	
			(0.107)	
Living on his own			0.355	
			(0.189)	
Community size (large city 100,000+ reference)				
Smaller city (25,000 – 100,000)				-0.201
				(0.103)
Suburb of a large city				-0.046
				(0.192)
Town less than 25,000				0.025
				(0.077)
Rural, non-farm				0.147
				(0.128)
Rural, farm				-0.132
				(0.072)
Constant	-6.879**	-6.823**	-6.931**	-6.795**
	(0.906)	(0.908)	(0.907)	(0.910)
N	4972	4972	4962	4964

[1] Standard errors in parentheses
[2] * significant at 5 %; ** significant at 1%

Table 8: Effect of Early Life Characteristics on Men's Mortality, National Longitudinal Survey of Older Men ($1966-1990$)

	Sample C (up to 1983)			
	Model 1	Model 2	Model 3	Model 4
Age	0.058**	0.058**	0.060**	0.059**
	(0.020)	(0.020)	(0.020)	(0.020)
Black	0.279**	0.258**	0.265**	0.299**
	(0.067)	(0.068)	(0.069)	(0.068)
Birth cohort ($1906-1910$ reference)				
$1911-1915$	-0.114	-0.114	-0.113	-0.112
	(0.118)	(0.118)	(0.118)	(0.118)
$1916-1921$	-0.357	-0.356	-0.344	-0.357
	(0.207)	(0.207)	(0.207)	(0.207)
R foreign-born	-0.509**	-0.512**	-0.502**	-0.525**
	(0.178)	(0.179)	(0.178)	(0.179)
Education of head of household (≤ 6 years reference)				
Missing		0.027		
		(0.076)		
$7-8$ years		-0.085		
		(0.099)		
$9-12$ years		0.046		
		(0.117)		
$13-$ years		-0.201		
		(0.165)		
Family structure (mother and father reference)				
Father and stepmother			-0.097	
			(0.253)	
Mother and stepfather			0.307	
			(0.190)	
Father only			0.241	
			(0.156)	
Mother only			-0.096	
			(0.111)	
Male relative			0.303*	
			(0.142)	
Other arrangement			0.055	
			(0.133)	
Living on his own			0.355	
			(0.239)	
Community size (large city $100,000+$ reference)				
Smaller city ($25,000-100,000$)				-0.135
				(0.130)
Suburb of a large city				0.065
				(0.236)
Town less than $25,000$				0.154
				(0.097)
Rural, non-farm				0.294
				(0.155)
Rural, farm				-0.026
				(0.092)
Constant	-7.079**	-7.035**	-7.180**	-7.136**
	(1.122)	(1.125)	(1.123)	(1.125)
N	4326	4326	4317	4321

[1] Standard errors in parentheses
[2] * significant at 5 %; ** significant at 1%

Table 9: Effect of Early Life Characteristics on Men's Mortality, National Longitudinal Survey of Older Men (1966 − 1990)

	Sample A (up to 1990)			
	Model 1	Model 2	Model 3	Model 4
Age	0.070**	0.070**	0.071**	0.070**
	(0.014)	(0.014)	(0.014)	(0.014)
Black	0.119*	0.107*	0.113*	0.135**
	(0.049)	(0.050)	(0.051)	(0.050)
Cohort (1906 − 1910 reference)				
1911 − 1915	-0.005	-0.004	0.003	-0.005
	(0.085)	(0.085)	(0.085)	(0.085)
1916 − 1921	-0.018	-0.017	-0.007	-0.021
	(0.148)	(0.148)	(0.148)	(0.148)
R foreign-born	-0.499**	-0.501**	-0.506**	-0.515**
	(0.121)	(0.121)	(0.122)	(0.121)
Education of head of household ($\leq$ 6 years reference)				
Missing		0.006		
		(0.055)		
7 − 8 years		-0.054		
		(0.071)		
9 − 12 years		0.009		
		(0.086)		
13− years		-0.098		
		(0.114)		
Family structure (mother and father reference)				
Father and stepmother			-0.018	
			(0.174)	
Mother and stepfather			0.377**	
			(0.135)	
Father only			0.166	
			(0.118)	
Mother only			0.012	
			(0.076)	
Male relative			0.138	
			(0.113)	
Other arrangement			0.029	
			(0.097)	
Living on his own			-0.056	
			(0.202)	
Community size (large city 100,000+ reference)				
Smaller city (25,000 − 100,000)				0.064
				(0.089)
Suburb of a large city				0.169
				(0.164)
Town less than 25,000				0.187**
				(0.070)
Rural, non-farm				0.198
				(0.119)
Rural, farm				0.022
				(0.066)
Constant	-7.561**	-7.532**	-7.633**	-7.648**
	(0.804)	(0.805)	(0.805)	(0.808)
N	4972	4972	4962	4964

[1] Standard errors in parentheses
[2] * significant at 5 %; ** significant at 1%

Table 10: Effect of Early Life Characteristics on Men's Mortality, National Longitudinal Survey of Older Men (1966 – 1990)

	Sample B (up to 1990)			
	Model 1	Model 2	Model 3	Model 4
Age	0.065**	0.063**	0.066**	0.064**
	(0.013)	(0.013)	(0.013)	(0.013)
Black	0.281**	0.262**	0.271**	0.302**
	(0.044)	(0.046)	(0.046)	(0.045)
Birth cohort (1906 – 1910 reference)				
1911 – 1915	-0.049	-0.051	-0.041	-0.051
	(0.078)	(0.078)	(0.078)	(0.078)
1916 – 1921	-0.111	-0.114	-0.098	-0.122
	(0.135)	(0.135)	(0.135)	(0.135)
R foreign-born	-0.166	-0.166	-0.167	-0.188
	(0.099)	(0.099)	(0.100)	(0.100)
Education of head of household ($\leq$ 6 years reference)				
Missing		0.051		
		(0.051)		
7 – 8 years		-0.057		
		(0.066)		
9 – 12 years		0.000		
		(0.079)		
13– years		-0.090		
		(0.105)		
Family structure (mother and father reference)				
Father and stepmother			-0.101	
			(0.164)	
Mother and stepfather			0.360**	
			(0.128)	
Father only			0.083	
			(0.111)	
Mother only			0.059	
			(0.069)	
Male relative			0.090	
			(0.105)	
Other arrangement			0.056	
			(0.087)	
Living on his own			0.161	
			(0.171)	
Community size (large city 100,000+ reference)				
Smaller city (25,000 – 100,000)				-0.043
				(0.081)
Suburb of a large city				0.150
				(0.147)
Town less than 25,000				0.085
				(0.064)
Rural, non-farm				0.137
				(0.108)
Rural, farm				-0.062
				(0.060)
Constant	-7.064**	-7.008**	-7.158**	-7.047**
	(0.735)	(0.737)	(0.736)	(0.739)
N	4972	4972	4962	4964

[1] Standard errors in parentheses
[2] * significant at 5 %; ** significant at 1%

Table 11: Effect of Early Life Characteristics on Men's Mortality, National Longitudinal Survey of Older Men (1966 – 1990)

	Sample C (up to 1990)			
	Model 1	Model 2	Model 3	Model 4
Age	0.072**	0.071**	0.073**	0.072**
	(0.015)	(0.015)	(0.015)	(0.015)
Black	0.228**	0.208**	0.214**	0.245**
	(0.051)	(0.053)	(0.053)	(0.052)
Birth cohort (1906 – 1910 reference)				
1911 – 1915	-0.043	-0.042	-0.034	-0.042
	(0.090)	(0.090)	(0.090)	(0.090)
1916 – 1921	-0.121	-0.12	-0.105	-0.124
	(0.157)	(0.157)	(0.157)	(0.157)
R foreign-born	-0.460**	-0.464**	-0.470**	-0.482**
	(0.130)	(0.130)	(0.131)	(0.131)
Education of head of household ($\leq$ 6 years reference)				
Missing		0.014		
		(0.058)		
7 – 8 years		-0.079		
		(0.074)		
9 – 12 years		-0.004		
		(0.090)		
13– years		-0.156		
		(0.121)		
Family structure (mother and father reference)				
Father and stepmother			-0.101	
			(0.191)	
Mother and stepfather			0.412**	
			(0.140)	
Father only			0.184	
			(0.122)	
Mother only			0.027	
			(0.080)	
Male relative			0.149	
			(0.117)	
Other arrangement			0.076	
			(0.101)	
Living on his own			0.087	
			(0.211)	
Community size (large city 100,000+ reference)				
Smaller city (25,000 – 100,000)				0.021
				(0.095)
Suburb of a large city				0.256
				(0.167)
Town less than 25,000				0.157*
				(0.074)
Rural, non-farm				0.203
				(0.124)
Rural, farm				0.001
				(0.070)
Constant	-7.573**	-7.525**	-7.702**	-7.648**
	(0.853)	(0.854)	(0.853)	(0.856)
N	4326	4326	4317	4321

[1] Standard errors in parentheses
[2] * significant at 5 %; ** significant at 1%