Towards a New, Public Data Set for Studying Mortality Inequality: Matching the 1940 U.S. Census with Social Security death records, 1963-2011

Joshua R. Goldstein (UCB)
Monica Alexander (UCB)

Extended Abstract for 2017 PAA Proposal

September 30, 2016

Abstract

The use of large, confidential administrative data sets has blossomed in recent years. But there is a dearth of public-use micro data. This makes replicability hard. It also slows advances in important issues such as the treatment of censored observations, modeling of covariates and frailty, and comparability with the general population. In this paper, we describe our first attempts to create a public use, microdata set for the study of mortality. We first link the publicly accessible 1940 U.S. census to the 2011 Social Security Death Index, achieving a high-certainty match rate of about 20 percent. We then develop techniques for estimating mortality rates from this linked data set, using extinct cohort methods, and compare the estimated death rates to the Human Mortality Database. We conclude with a discussion of possible approaches to making this linked data set appropriate for public use.

1 Background

In the field of mortality research, the use of large, confidential administrative data sets has blossomed in recent years. Waldron (2007) used Social

Security earnings and mortality micro-data. More recently, Chetty et al. (2016) and Saez and Zucman (2014) have used IRS micro-data. These studies find large – and perhaps steepening – gradients in mortality by social status, income, and geography. These individual researchers have gained access to these confidentiality-protected, in-house data sets. However, unlike the public use surveys created by government and the research community, administrative data remains largely inaccessible for the general research public. These studies are therefore difficult to replicate and extend. Basic issues like the treatment of censored observations, the role of selective mortality and heterogeneity, the measurement of covariates, the temporal comparability of inequality measures, and the generalizability of the research to the broader population are all difficult to assess and address.

Inspired by the success of the Human Mortality Database (HMD) at Berkeley and the Max Planck Society and the Integrated Public Use Microsamples (IPUMS) at the University of Minnesota, the goal of this project is to create large-scale, public micro-data with detailed information on individual level covariates and on mortality.

In this paper, we show that a large, rich, and representative data set can be constructed from publicly available sources that have recently become available. Notably, the release of the 1940 U.S. Census provides us with a population of 130 million individuals that can potentially be followed over time.[1] The Social Security Death Index provides individual death records for approximately 80 million people from the mid-1960s to 2011 (or 2013). These two data sets can be matched using first names, last names, and year of birth. The resulting data set then contains the complete census information, which is unusually rich in 1940, along with the date of death for those individuals who are matched.

In this paper, we describe the matched data set and carry out several analyses on data quality. Specifically, we aim to:

1. Analyze the selectivity of individuals who are matched.

[1] Public access to individual records is allowed by law beginning 70 years after each census. Images for the 1940 census are all available. The 1940 census was transcribed by a joint-venture of the University of Minnesota Population Center and ancestry.com. The complete 1940 census without names is available at IPUMS USA. The complete 1940 census with names is available through a secure protocol. We thank MPC and ancestry.com for allowing the UC Berkeley Department of Demography to host the complete census counts we used for this project.

2. Compare the implied mortality rates with national cohort mortality rates from the Human Mortality Database

   and

3. Show that the resulting data can be useful in quantifying inequalities in mortality.

2 The Data Sets

In the United States, census micro data is made publicly available 70 years after each census. Since the release of the images of the 1940 census forms, they have already been fully digitized. The full nominal transcription, with first and last names, is available by special agreement with ancestry.com and the University of Minnesota Population Center. The names and ages in this data set provide enough information to link a substantial fraction of census respondents to individual death records.

The 1940 census is not only the most recent publicly available census, it is also one of the richest in terms of individual information that was ever collected. For the first time, respondents were asked about their income and years of educational attainment. Information on the value of the house was asked of house owners, and monthly rent was asked of renters. The timing of the 1940 census was before the war-time mobilization that soon followed. After many years of the Great Depression, 1940 gives us a chance to observe the population in something approach "normal times."

For death records, we rely on the publicly available Social Security Death Masterfile (SSDM).[2] This file contains full names, exact dates of birth and exact dates of death for about 80 million deaths from 1963 to 2011. The work reported here uses the publicly available download. (We plan to obtain formal permissions from the e Social Security Administration to use this publicly available data.)

We estimate that about 95 percent of all deaths from 1975 to 2005 are included in the SSDM files.

We note that the SSDM files also contain Social Security numbers for all deaths. This potentially allows for further linkage to lifetime earnings reports. A number of projects have are being developed to provide individual identification numbers for every person in the 1940 census. These projects

[2] Copies are available for download at the activist website, `sssdm.info` .

will be using the individual identification numbers to match administrative records, and if successful will provide researchers with the means to work with protected data in a secure setting. One can view the work here as an early example of the kind of work that will be done in the decade to come, with the added benefit that the linked deaths can be made publicly available without the need to create a secure data environment.

We anticipate that we will be able to release the data in two ways. First, we will be able to release a file without any nominal identifying information that will have an identifying serial number that can be matched to the publicly available IPUMS microdata. This file will have date-of-death for those individuals that could be matched. In essence, this will allow the user to add one variable, date of death, to the existing 1940 census files. Second, this same file can be used by researchers with access to the nominal data.

3 Matching algorithm and match rates

Our approach to matching aims to create matches with high certainty and to minimize selectivity in matching, so that our matched data set is generalizable to the population.

The method we use is to create a key based on the first name, last name, and census age. First name is defined as the first word given in the first name field. Thus "John F." and "John Fitzgerald" both have the first name "John". "J. Fitzgerald" is coded as "J." Census age is defined as the age report in the 1940 census and as the age as of April 1 based on the exact birth date given in the Social Security data.

As a preliminary analyses, we matched the 1940 census records from California for males aged 20 to 60 in 1940 to the Social Security Death Records.[3]

After removing names that were incompletely transcribed and those for whom age was not given, we began with a candidate population of 2.2 million men aged 20 to 60 in California. After restricting our candidates to be matched to those with unique keys – that is, unique combinations of first

[3] To ease the preliminary analysis, we restricted ourselves to death records of people with California-given social security numbers. This geographic restriction will not be applied in the full analysis. When we no longer condition on geography, we will reduce the number of unique keys. However, we will also be able to match those who had Social Security numbers issued in a different state than their 1940 census residence.

name, last name, and age in census, we arrived at a candidate pool of 1.9 million men that could be matched. Matches were found found for about 440 thousand of these men, a match rate of about 23 percent.

There are three primary reasons for non-matching. These are:

1. The individuals did not die in the period from 1975 to 2005, either dying before or after.
2. The individuals died but were not included in the SSDM, most likely because they were not registered with Social Security. This group is small, however, since the total number of deaths for those 65 and over in the SSDM is about 95 percent of the total from Vital Statistics (available through the Human Mortality Database).
3. Individuals had inconsistent names and/or ages in the two data sources. We matched only when the match criteria were exactly met. Thus "E. Tufte" would not match "Ed Tufte", and neither would match "Edward Tufte".

```
California Male Population 1940, aged 20-60:  2.2 million
(with unique keys) : 1.9 million
(matched to death records): 440 thousand
```

For the paper we will quantify each source of non-matching. In particular, it will be possible to quantify what share of cohort deaths would be observed based on published cohort mortality rates in the HMD.

4 Mortality estimation and comparison with the Human Mortality Database

For each cohort, we observed a distribution of deaths that is truncated both on the left and right. For example, restricting deaths to the period 1975 to 2005, when the records are fairly complete, gives deaths aged 65 to 95 for the cohort of 1910.

Our approach to estimating mortality is to use reverse survival methods applied to extinct cohorts, in which the number of individuals in a cohort at risk at each age are the sum of people dying at that age or above. If the cohort is not yet extinct (right-censored), then the number of people at risk will be under-estimated using the reverse survival method. In order to solve this problem, we use the Human Mortality Database to estimate the fraction of each cohort surviving in 2005 and incorporate this into our estimation.

The figure below shows the estimated age-specific mortality rates for the cohorts aged 20 to 60 in 1940 (born 1880 to 1920). We see that we obtain very good correspondence between the Human Mortality Database estimates for the national population and our estimates from the matched data for California. We see that adjusting for survivorship after the last observed age allows us to obtain age-specific mortality rates for each cohort that track HMD estimates very closely.

Questions that remain to be resolved are: (1) validating that the unbiasedness of our mortality rates also applies when we disaggregate by income, education, and other covariates (2) adjusting for the missing deaths at older ages in the analysis of micro data.

5 Determinants of mortality, preliminary results

We provide two illustrations of how this micro-data can be used with covariates. First, we estimate hazards for those above and below median income (following the example of Waldron). The figure below shows that lower earners have higher mortality at younger ages but there appears to be convergence at older ages. We note that this analyses has not adjusted for right-censoring, but for this cohort, this effect should be small.

Second, we regressed age-at-death on selected covariates. We use a linear model, a simple multivariate regression on age-at-death. Typically, one would use event history (survival analysis) techniques to model hazards rather than simply modeling age at death, but since we have no right-censoring of the observations that we are able to match, this non-conventional approach may produce reasonable estimates. (We emphasize that the purpose of showing these results is just to give an idea of what can be done. We are not endorsing this particular statistical technique.)

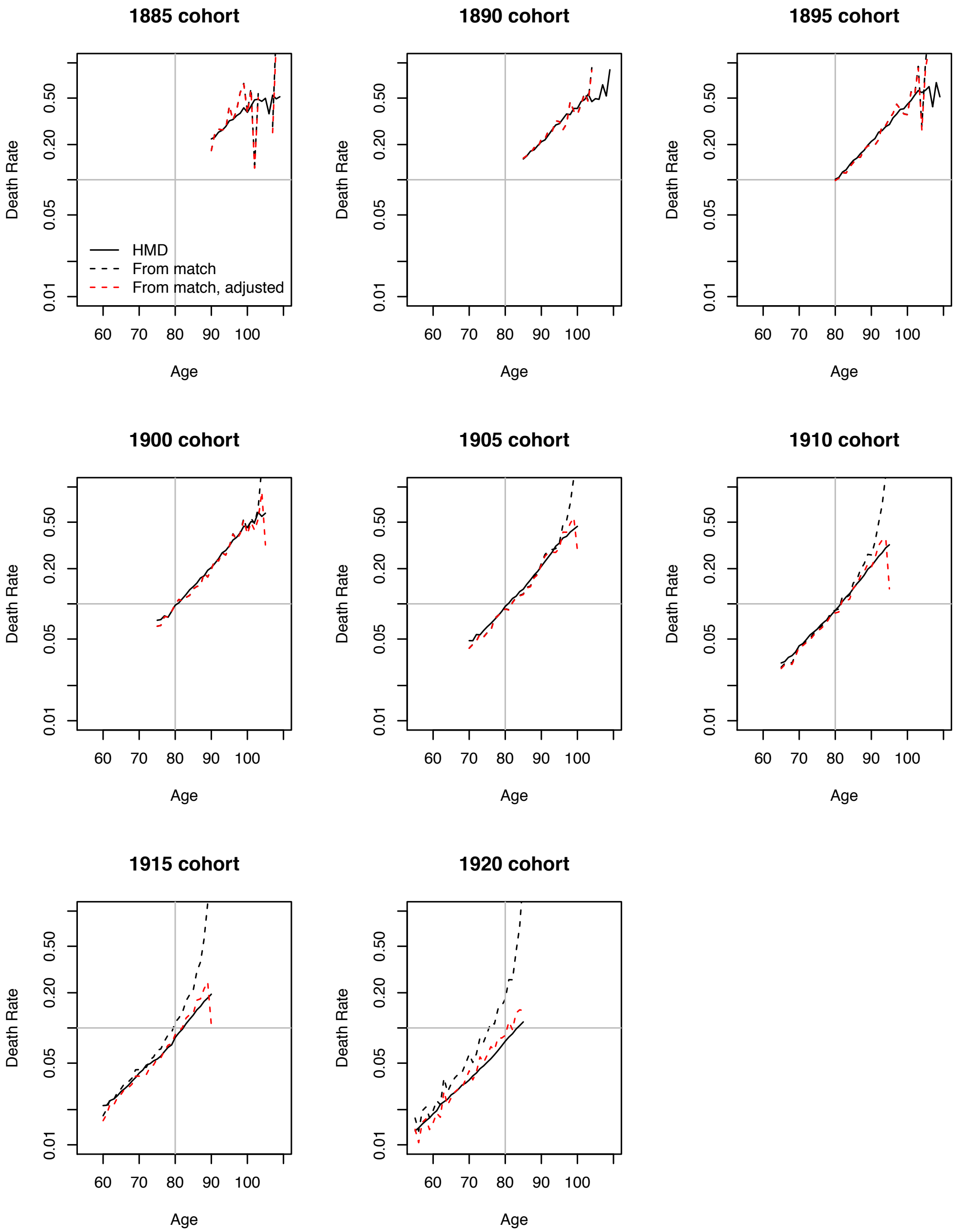

Figure 1: Age-specific hazard rates by cohort estimated from matched data from the 1940 Census and Social Security death records. Comparisons are with the Human Mortality Database. Adjusted hazards correct for cohort survivors after 2005. Cross-hairs allow notice of declines in mortality over time.

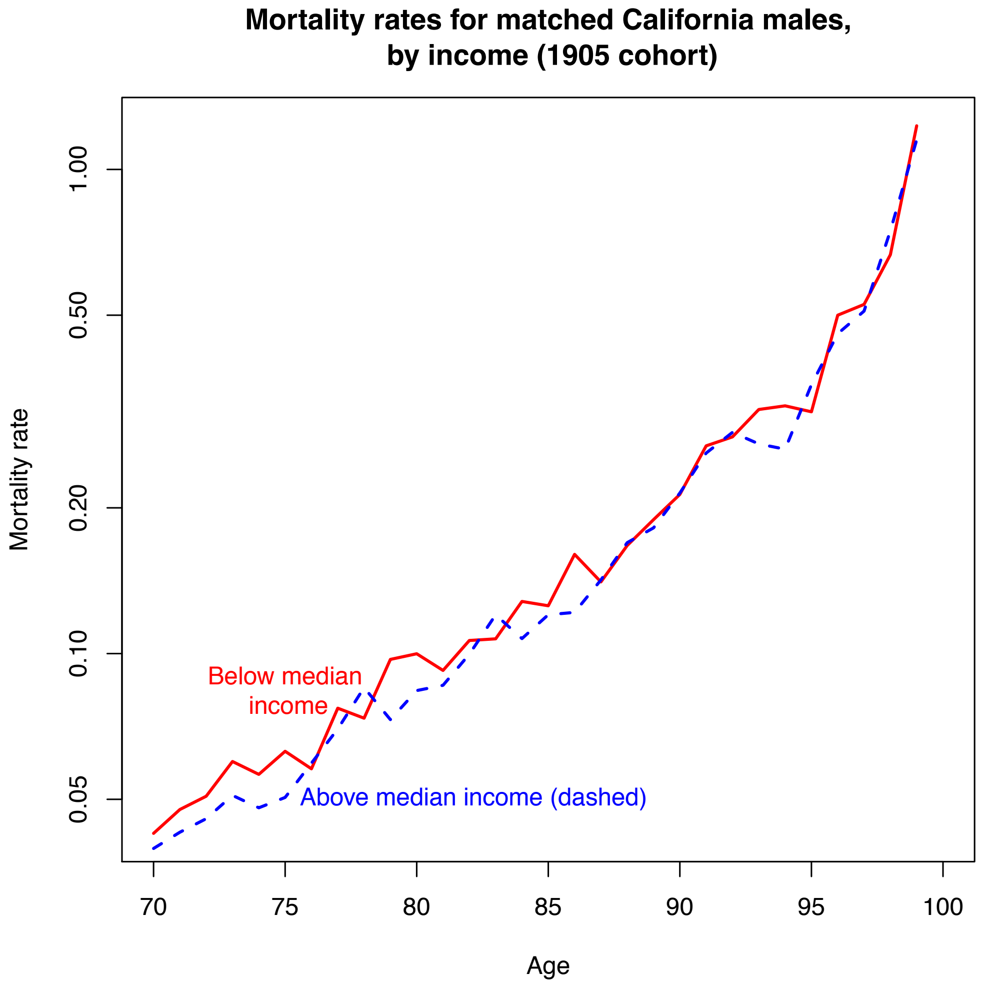

Figure 2: Age-specific hazard rates from 1940 Census-Social Security matched data for the California cohort of 1905 for male household heads, stratified by above and below median income. The lower earners have higher hazards up to about age 85, after which there is convergence.

```
====================================================================================================
                             m.race     m.income     m.educ     m.urban    m.owner    m.all
----------------------------------------------------------------------------------------------------
 (Intercept)                71.7***     67.7***     69.8***     71.7***     72.5***   68.8***
 age in 1940 (dummies)      yes         yes         yes         yes         yes       yes
 racesimp: black/white      -0.3                                                       0.2
 racesimp: chinese/white     1.4**                                                     2.2***
 racesimp: filipino/white    3.3***                                                    4.2***
 racesimp: japanese/white    2.7***                                                    3.0***
 racesimp: other/white      -1.9**                                                    -1.3*
 log(income)                             0.6***                                        0.3***
 educ                                                0.2***                            0.2***
 urban: urban/rural                                             -0.0                  -0.2***
 ownership: rent                                                            -0.9***   -0.8***
----------------------------------------------------------------------------------------------------
 R-squared                  0.083       0.085       0.086       0.083     0.085       0.088
 N                         250105      250105      247375      250105     249908      247180
====================================================================================================
```

Table 2: OLS regressions on age at death using Census-Social Security linked data for California Males

We include fixed effects (dummies) for age in the 1940 census. These are necessary because each cohort is differentially left-censored. This approach also allows for income and education effects to be estimated within cohorts, even though we estimate only one coefficient for income and for education.

The interpretation of our results is that a 10 percent increase in income (in 1940) would increase age at death by $0.10 \times \beta_{log(income)} = 0.06$ years.

An additional year of education would increase age at death by $\beta_{educ} = 0.2$ years.

Interestingly, Blacks do not have lower survival than Whites. But immigrant groups (the Chinese and Japanese and Filipinos) all live longer.

Home-owners (in 1940) live 0.9 years longer than renters (in 1940), a remarkably large effect, particularly since it persists even after controlling for income and education.

The final column includes all of these variables together, and we see among other things that the effect of education persists when income is controlled for, but that the effect of income is halved by controlling for education.

We note that in 1940, California had a population of about 7 million, or about $7/130 = 5$ percent of the U.S. population. So the sample size resulting from the complete national data set will be about 20× the size of this data set. This will allow detailed study of how different variables influence mortality in different ways in different contexts. It will also allow a detailed study of heterogeneity-based mortality selection, noted by Waldron as one possible explanation for the divergence of mortality in recent years.

The great limitation of this project is that in its present form it only includes measurement of covariates in 1940. This simplifies any analysis, but also means that measures in 1940 will not perfectly represent the life-time experience of the individuals. However, we see this work as complementary to other efforts to link the 1940 to longitudinal data sources on economic and social measures.

6 Discussion

We tentatively offer the following concluding observations

- Conservative linking can produce large data files that as a whole give mortality representative of the large population.

- Characteristics observed in 1940 are predictive of age at death.

- Existing demographic techniques can be used to estimate mortality rates for the limited time-window of deaths we see for each cohort.
- All of this can be done with existing public data and the linked files can be shared without violating privacy protections.

We hope that the linking of death records to the 1940 census will provide a valuable, public use data set for mortality researchers.