Berkeley Unified Numident Mortality Database: Public Administrative Records for Individual Level Mortality Research

Joshua Goldstein*
Casey Breen†

06 March, 2020

Abstract

With the release of Social Security application (SS-5), claim, and death records, the National Archives and Records Administration (NARA) has created a new administrative data resource for researchers studying mortality. While much progress has been made in understanding the demographic determinants of mortality in the United States using survey data, the lack of a population-level register data is a barrier to further advances in mortality research. This publicly available micro-level dataset provides researchers access to over 49 million mortality records with demographic covariates and fine geographic detail, allowing for high-resolution mortality research. In this paper, we document the contents of this dataset, provide access to a cleaned and harmonized version of the data, and discuss statistical methods for estimating mortality differentials based on this deaths-only dataset.

Introduction

The Numerical Identification System (Numident) forms the backbone of the U.S. Social Security Administration's record keeping system. For every person with a Social Security number, the Numident tracks claims status, date of birth (and, if applicable, death), as well as other background information including birthplace, race, sex, and names of parents. In 2013, the Social Security Administration transferred a large portion of their Numident records to the National Archives and Records Administration (NARA). The public released these records in 2019, which we call "NARA Numident", offers nearly complete coverage of those who died from 1988 to 2005. In this paper, we describe the contents of the publicly available

*josh.goldstein@berkeley.edu; Department of Demography, UC Berkeley
†caseybreen@berkeley.edu; Department of Demography, UC Berkeley

records, introduce a cleaned and harmonized version of the data, and show how the records can be used for the study of mortality in the United States.

The NARA Numident is individual-level dataset covering over 49 million people who have died. In addition to individual identifiers, NARA Numident includes information on race, sex, birthplace, ZIP Code of residence at the time of death, and administrative variables, such as a person's age when they submitted their first Social Security application and their total number of Social Security applications. There are no direct measures of socioeconomic status in the NARA Numident. To overcome this, the records can be linked using such individual identifiers to obtain income and education covariates (Goldstein et al. 2019). Identifiers include Social Security numbers as well as names, birth place, and dates of birth. The death coverage is nearly complete for deaths to persons age 65+ for the window of 1988-2005.

The public release of the NARA Numident makes it one of the first administrative data sets on mortality that can be used by all researchers. Previously, Social Security Numident records have been used to study mortality by researchers employed by the Social Security Administration (e.g. (Waldron 2007)) or collaborating with SSA researchers (e.g (Mehta et al. 2016)). Researchers using restricted access IRS data have also carried out mortality research (Chetty et al). Our hope is that the public availability of this data will encourage more mortality research using administrative records, enhance the replicability and debate about results, and open up new avenues of research.

Administrative data offers several advantages for the study of mortality. The large sample sizes enable the comparison of ages, birth cohorts, small sub-populations, and small geographic areas. The large sample sizes also enable the study of mortality at the oldest ages, when their are only few survivors. An additional advantage is that the public nature of the NARA Numident means that individual identifiers can be used to link to other data with covariates of mortality. Since there are no restrictions to the use of this public data, researchers can also link records to their own restricted data sets.

The NARA Numident records pose a challenge for mortality estimation. Because the data set includes only those who have died, there is no measure of survivorship. Traditional statistical methods relying on exposure to risk are not appropriate. Instead, one needs to use methods that rely on the distribution of deaths by age within cohorts. We discuss these methods below and provide examples of their use.

Note: next paragraph is redundant – need to check to see if all info is already in above text] While administrative mortality datasets have been used by a small set of researchers who have been able to work with government employees inside restricted computing environments (Chetty et al. 2016; Mehta et al. 2016), the NARA Numident data is openly accessible to all researchers. Our hope is that the public availability of this data will encourage more mortality research using administrative records, enhance the replicability and debate about results, and open up new avenues of research. To facilitate mortality research with the NARA Numident records, we have created a cleaned and harmonized version of the Numident records with enhanced documentation: the Berkeley Unified Numident Mortality Database (BUNMD).

The methods we provide here are also useful for researchers working with the Social Security Death Master File (DMF), another publicly available data resource for mortality research. The DMF was first made available in 1988 and is extracted quarterly from the Numident (Hill and Rosenwaike 2001). The file has been used by some researchers to study mortality, particularly at older ages (Gavrilov and Gavrilova 2012). While the DMF has high death coverage for the wider window of 1975 to 2005, it lacks most of the covariates available in the NARA Numident records (Hill and Rosenwaike 2001).

We are also in the process of linking both the DMF and the NARA Numident records to the full-count 1940 Census, to create a rich, publicly linked administrative dataset for the study of mortality (Goldstein et al. 2019).

The Structure and Content of the NARA Numident files

The NARA Numident records contain three types of entries: applications, claims, and deaths. The Social Security Administration adds a new entry to the Numident when a Social Security cardholder submits a new application or claim. New entries never overlay old entries. Instead, a new entry is added to the pre-existing Numident, ensuring that information is never overwritten. Figure 1 shows the distribution of application and claim entries per person. In the NARA Numident records, 43.3% of persons have multiple application entries, 0.3% of persons have multiple claim entries, and 0% have multiple death records.

To illustrate the structure and content of the NARA Numident records, we show the released records for the actress Lana Turner, who died in 1995, and for the Supreme Court Justice Thurgood Marshall, who died in 1993. For Thurgood Marshall, we have one application and one death record. For Lana Turner, we have one death record and four application records, likely corresponding to name changes each time she was married.

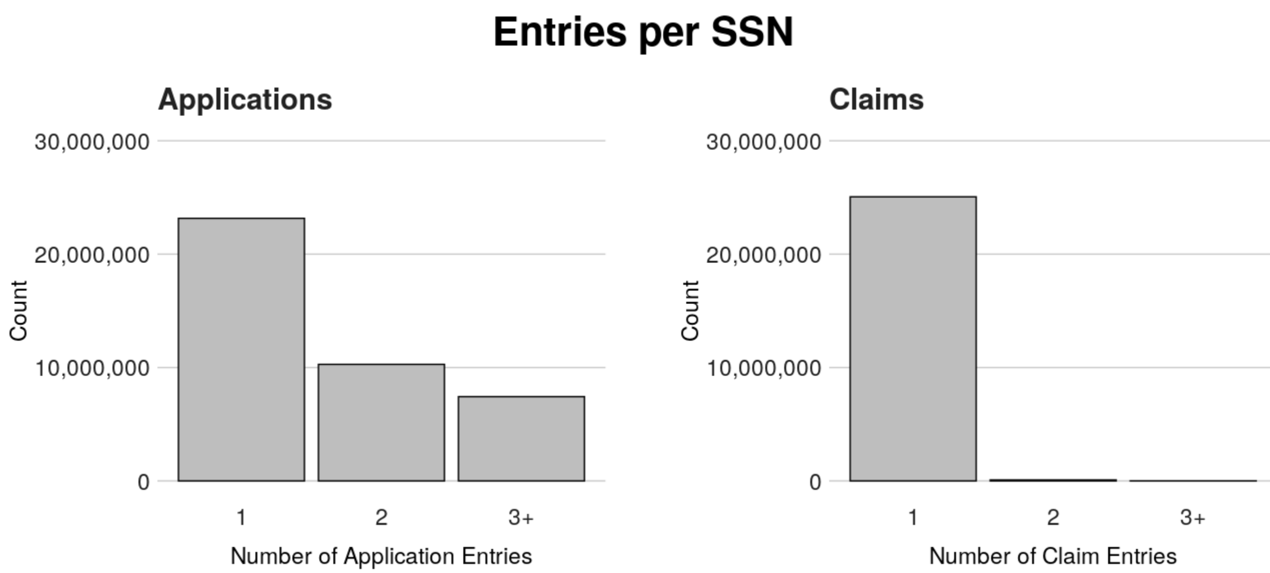

Figure 1: Number of entries per person for the Numident Application and Claims files.

Table 1: Constructing the BUNMD from NARA Numident Records

Thurgood Marshall

	ssn	fname	lname	birth date	sex	race
Application Entry 1	131074264	THURGOOD	MARSHALL	7/2/1908	1	**2**

	ssn	fname	lname	birth date	death date	sex
Death Entry	**131074264**	**THURGOOD**	**MARSHALL**	**7/2/1908**	**1/24/1993**	**1**

	ssn	fname	lname	byear	bmonth	bday	dyear	dmonth	dday	death_age	sex	race_first
BUNMD Entry	**131074264**	**THURGOOD**	**MARSHALL**	**1908**	**7**	**2**	**1993**	**1**	**24**	**84***	**1**	**–**

Lana Turner

	ssn	fname	lname	birth date	race	sex
Application Entry 1	567183907	LANA	TURNER	2/8/1921	**1**	2
Application Entry 2	567183907	LANA	TOPPING	2/8/1921	1	2
Application Entry 3	567183907	LANA	BARKER	2/8/1921	**1**	2
Application Entry 4	567183907	LANA	DANTE	2/8/1921	–	2

	ssn	fname	lname	birth date	death death	sex
Death Entry	**567183907**	**LANA**	**TURNER**	**2/8/1921**	**6/29/1995**	**2**

	ssn	fname	lname	byear	bmonth	bday	dyear	dmonth	dday	death_age	sex	race_first
BUNMD Entry	**567183907**	**LANA**	**TURNER**	**1921**	**2**	**8**	**1995**	**6**	**29**	**74***	**2**	**1**

Note : Bolded values were selected for in the BUNMD. Starred values were represent contructed variables.

	bpl		
	MD		

		zip_residence	
		220411335	

race_last	bpl	zip_residence	number_apps
2	**2400**	**220411335**	**1***

	bpl		
	ID		
	ID		
	_		
	_		

		zip_residence	
		900255240	

race_last	bpl	zip_residence	number_apps
1	**1600**	**900255240**	**4***

We introduce a cleaned and harmonized version of the NARA Numident records: the Berkeley Unified Mortality Numident Database (BUNMD). This file condenses the Numident death, application, and claims records into a single file with one record per person. This file is available for download at _____ . The file includes about 49 million records, 28 variables, and is about 5.7 Gb in size. For Lana Turner, the BUNMD data record is:

The BUNMD condenses the NARA Numident records into a single file with one record per person. The original NARA release contained 49,459,293 death records entries, 72,120,516 applications entries for 40,870,455 unique persons, and 25,228,257 claims entries for 25,140,847 unique persons. To construct the BUNMD, we first selected key variables from the death records. For each record with a death entry, we added additional covariates from the application and claims entries. For individuals with multiple application or claims entries, we used a set of decision rules to reconcile discrepant values across entries (see technical appendix for more details). Finally, we constructed variables reporting (1) total number of applications, (2) total number of claims, (3) age at first Social Security application, (4) state in which the Social Security number was issued. Figure 2 shows the process for constructing the BUNMD. In order to study name changes, race changes, and other features, the original NARA Numident records are useful, and are available upon request.

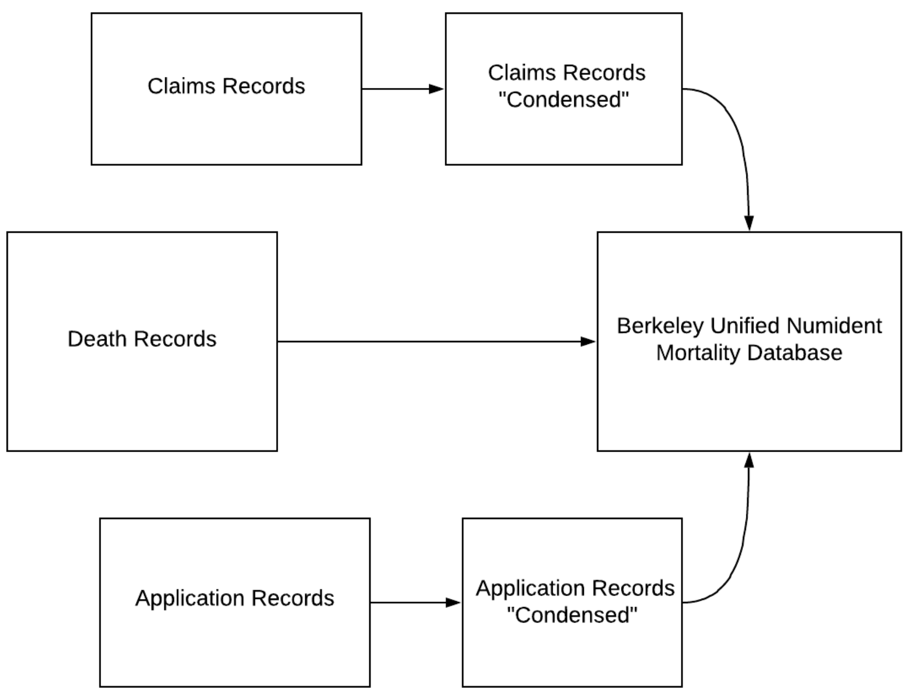

Figure 2: Berkeley Unified Numident Mortality Database creation flowchart.

Numident Coverage

The NARA Numident records are a subset of the complete Numident. One challenge of working with the NARA Numident records is that the Social Security Administration's process for selecting the records to transfer to NARA for their public release is not well-defined. The NARA documentation states that the first transfer of records contained: "individuals with a verified death between 1936 and 2007 or who would have been over 110 years old by December 31, 2007" (47 2019). This alone, however, does not clarify the patterns of death coverage in the NARA Numident. Figure 3 compares the total number of deaths for persons age 65+ in the BUNMD to the Human Mortality Database (HMD). Death Coverage is nearly complete between 1988 and 2005. Figure 4 shows the coverage visualized on an age-period Lexis surface, an established demographic visualization technique (Schöley and Willekens 2017). Each cell represents death coverage, measured as the ratio of the total count of deaths in the BUNMD to the total count of deaths in HMD for a given age and year.

We create two BUNMD samples with high death coverage. Sample 1 includes deaths to persons age 65+, occurring between 1988 to 2005, from the birth cohorts of 1900 to 1940. Sample 2 is the subset of Sample 1 records with complete information on sex, birthplace, and race. For each sample, we constructed inverse probability weights to the Human Mortality database on age at death, year of birth, year of death, and sex.

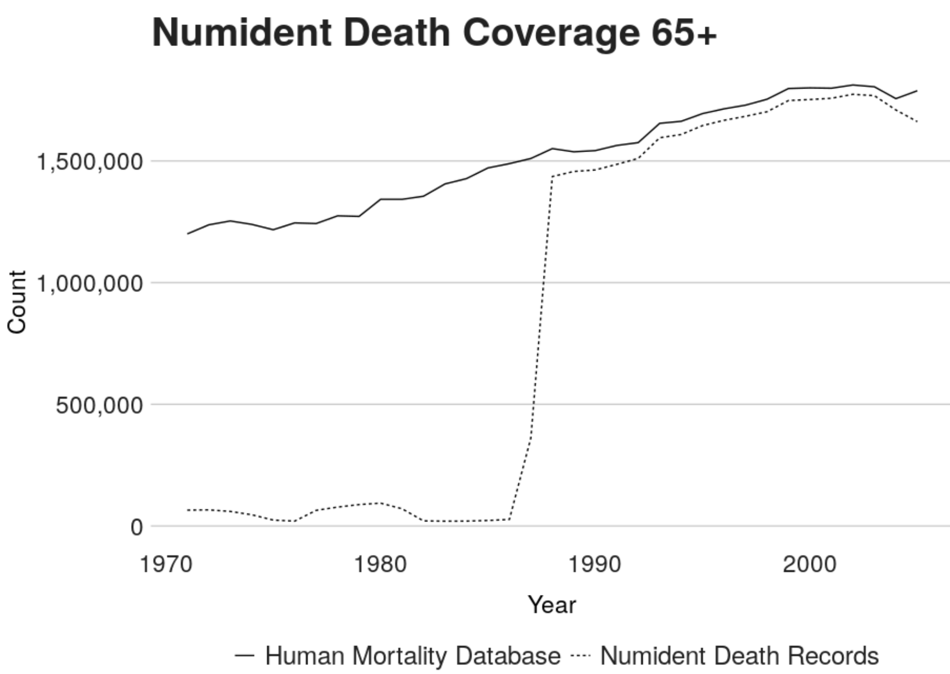

Figure 3: BUNMD Death Coverage for persons 65+

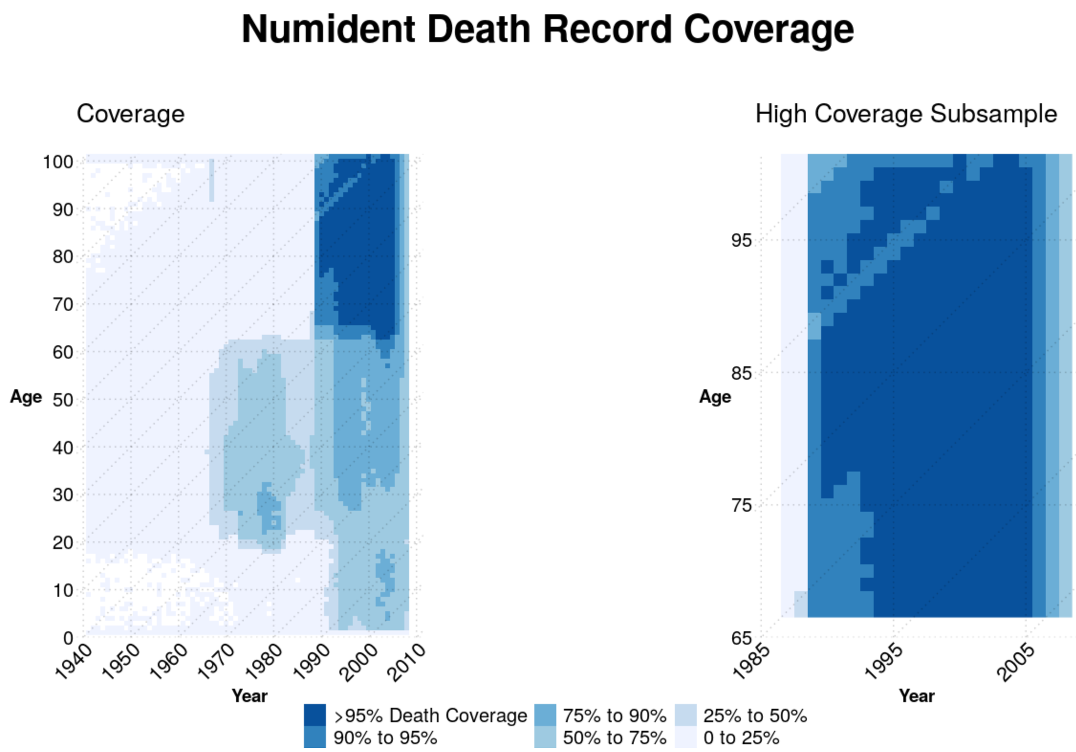

Figure 4: Lexis diagrams of BUNMD death coverage. The left panel shows the BUNMD death coverage for 1940 to 2010. The right panel zooms in on age-period area where coverage is highest.

Estimation: Deaths without Denominators

The BUNMD file includes only individuals who have died. For extinct cohorts (in which all members have died), it is possible to use classical methods of "extinct generations" to calculate mortality rates. These methods are appropriate for the cohorts born before 1900, for which only a few survivors to age 105 will die after 2005. For later cohorts, however, we have developed several different methods, which can be chosen based on suitability for the research question of interest.

The first method is to fit parametric survival models (Gompertz and Makeham), using maximum likelihood for doubly-truncated cohorts. The second method is to use ordinary regression, inflating the observed coefficients in order to account for truncation. Finally, we introduce the Cox regression method.

Method 1: Parametric survival models

Human mortality has a characteristic pattern in older ages. To a first approximation — first noticed by Benjamin Gompertz — mortality hazards rise exponentially with age.

$$h(x) = ae^{b*x} \tag{1}$$

The constant exponential rate of increase is most pronounced from ages in the 70s to ages in the 90s. At younger ages, between 40 and 70, mortality is often somewhat higher than would be predicted by a Gompertz model. This was first noted by Makeham, who suggested that adding a constant term would be a better description of observed mortality.

$$h(x) = c + a \cdot exp(b \cdot x) \tag{2}$$

Finally, at older ages, although there is still much debate, there may be a leveling of mortality. Thus the logistic model has been introduced to account for this leveling of mortality. For any parametric model, it is possible to write down the likelihood given the deaths we observe. For truncated cohorts, with known left-truncation L and known right truncation R, we can write down the likelihood as

$$L = \prod(f_i(\theta)/(F_R(\theta) - F_L(\theta))) \tag{3}$$

The estimates of the vector θ of parameters can be obtained by maximizing the likelihood, or, equivalently, the log-likelihood.

Method 2: Ordinary Least Squares Regression

Regression on age at death is an easy and effective way to analyze the Numident mortality data. Regression coefficients tell the effect of covariates on the mean age at death. Because left and right truncation ages vary by cohort, it is important to include fixed effect terms for each year of birth. Models of the form:

$$\text{Age\_at\_death} = \text{birth\_year\_dummy} + \text{covariates of interest} \tag{4}$$

provide estimates of the effect of the covariates on the age of death in the sample, controlling for birth cohort truncation effects.

Truncation, however, will tend to bias downward the estimated effects of any covariates (Greene 2005). Truncation excludes the tails of the distribution, thus reducing the average difference between groups. The idea is that the average differences between groups will be measured to be much smaller if we exclude the tails of the distribution.

Simulation tells us that the magnitudes of the regression coefficients need to be inflated by a factor of about 2 or 3 for many of the cohorts that are covered by the Numident files. The table below gives the inflation factors for each cohort, based on a simulation of a Gompertz distribution with M = 79.6 and b = 0.0826 (the values found by fitting to the untruncated cohort of 1910 using HMD data). The interpretation of these numbers is that a regression coefficient of 0.5, say comparing Men and Women, found using the data from the cohort of 1910 (observed from 1988 to 2005) translates to a difference of life expectancy at age 65 of 0.5 x 2.3 = 1.15 years.

Method 3: Cox regression for extinct cohorts

For cohorts that are extinct (or very nearly so), Cox regression provides a convenient method. Cohorts born in 1900 or earlier are observed to age 105. Cox regression makes no distributional assumptions about the shape of mortality, but does assume proportional effects on the hazards. (Wachter (2014)).

Case Studies

The old-age mortality of the foreign born

The mortality of immigrants is often lower than natives, despite the fact that many immigrants are often disadvantaged in terms of education and income. The "immigrant paradox" has long been observed for Mexican immigrants, one of the only immigrant groups of sufficient number to produce accurate mortality measures from sample surveys. Recently, (Mehta et al. 2016) were able to use internal Social Security and Medicare records, finding that a diverse set of immigrant groups had lower mortality than natives.

Here, we first show how the BUNMD data can be used to confirm Mehta's findings using publicly available data. We then take advantage of the information on race to look at variation within Cuban immigrants. There are many other topics that can be investigated relating to the mortality of immigrants, including spatial patterns based on zipcode of residence at the time of death and cultural variables that can be measured using first and last names (Goldstein and Stecklov 2016).

In our analysis, we restrict ourselves to foreign-born individuals who applied for Social Security cards before turning age 65 and before 1988. This assures that the distribution of deaths we observe is not biased upwards by immigrants arriving in the midst of our observation period.

For the study of race, we also restrict ourselves to individuals who recorded a race before 1980, when the only options were “White”, “Black”, and “Other”.

Country-of-birth differences

To measure mortality differences by country of birth, we have chosen the most 20 or so most common origins for immigrants in our sample who were born from 1910 to 1919. We fit a regression model separately for men and women, with fixed effects for year of birth. This approach is aimed at reducing compositional effects that stem from observing different ages of death for each birth cohort.

Figure shows the difference in mean age of death between natives and the foreign-born. Differences in mean ages in the truncated sample understate differences in life expectancy at age 65. A good approximation for translating the differences in the truncated sample to life expectancy at age 65 is to multiply the regression coefficients by about 4. We discuss how such multiplicative factors can be estimated in section .

The large sample size gives us enough precision to see that there are interesting patterns by individual country. Whereas Mehta reported results for broader regions and found that immigrants from every region had lower mortality than natives, the country-level estimates we report here show that one group, Irish men, suffers a mortality disadvantage relative to the native-born.

For both sexes, we see that the longest lived groups come from a remarkable variety of origins. A prominent explanation of immigrant mortality advantage is selective migration: those who overcome obstacles to migration are a select and quite healthy group. This theory has some support from the pattern we see, with those born in countries that are the farthest away, e.g., Greece, Philippines, and China all being among the longest lived, and those coming from relatively close, English speaking countries (Canada, England and Ireland) as having the smallest longevity advantage. The Mexican case is interesting in that it is an immediate geographic neighbor, but non-English speaking. Male migrants from Mexico are among the longest lived, but female migrants from Mexico do not appear to have a particularly large advantage over natives when compared to other countries of origin.

There are many interesting avenues of research on the mortality advantages of immigrants that could be carried with the Numident data. Geographic variation, residence in higher and lower income areas, residence in areas with other immigrants, differences by immigration cohort (as proxied by age of first application for Social Security), and racial and ethnic differences could all be pursued. In addition, first and last names can also be analyzed for indications of ethnic diversity within immigrant groups and for measures of acculturation (Goldstein and Stecklov 2016).

Racial differences among Cuban immigrants

We saw above that Cuban immigrants are among the longest living sub-groups in the United States. We can use the NARA Numident to ask explore further whether immigrants from this racially diverse country, differ in their longevity in the United States, keeping in mind that racial identity is self-identification on the Social Security SS-5 application. We restrict ourselves to pre-1980 responses to the race question, when the only options were “White” (N = 35,000), “Black” (N = 1,000) and “Other” (N = 800).

We report the results of a regression of age of death on race, sex, and birth year in the Table below. We can see that Black Cuban immigrants died earlier than “White” Cuban immigrants. The effect of -0.8 years in the regression corresponds to a life expectancy difference at age 65 of about 3 years. (There is no clear difference between Cuban immigrants that identified as “Other” and those who identified as “White”.) The disadvantage of Black Cuban immigrants is consistent with racial inequality in Cuba and with the reception of Cubans in the United States (Newby and Dowling 2007).

This analysis could be extended by looking at the residential patterns of black and white Cuban immigrants in the United States. There are also other immigrant origins, such as the Dominican Republic and Brazil, which are racially diverse. A deeper dive into the original SS-5 application files may also reveal interesting patterns about who chooses “Hispanic” identity and when. (**???**), using Numident records linked to the 1940 census for example, finds that those who change their identity from “Hispanic” to “White” tend to have higher earnings and educational attainment.

Table 1

	Dependent variable:
	death_age
Race_pre_1980: White/Other	0.284* (0.169)
Race_pre_1980: Black	−0.801*** (0.167)
Byear1900	6.631*** (0.887)
Byear1901	5.051*** (1.012)
Byear1902	4.243*** (0.543)
Byear1903	4.521*** (0.402)
Byear1904	4.519*** (0.309)
Byear1905	3.300*** (0.259)
Byear1906	2.467*** (0.198)
Byear1907	2.226*** (0.180)
Byear1908	1.280*** (0.148)
Byear1909	0.189 (0.128)
Byear1911	−0.431*** (0.110)
Byear1912	−1.163*** (0.116)
Byear1913	−1.920*** (0.118)
Byear1914	−2.649*** (0.122)
Byear1915	−3.431*** (0.131)
Byear1916	−4.257*** (0.134)
Byear1917	−5.150*** (0.134)
Byear1918	−5.859*** (0.131)
Byear1919	−6.782*** (0.133)
Byear1920	−7.786*** (0.133)
Sex: Female	1.809*** (0.051)
Constant	84.418*** (0.084)
Observations	36,952
R^2	0.282
Adjusted R^2	0.282
Residual Std. Error	6.722 (df = 36928)
F Statistic	630.775*** (df = 23; 36928)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Regression results for mean age at death of Cuban immigrants by race, birth year,

and sex.

Translating regression results into life expectancy at age 65

The regression results we show report differences in mean age at death observed in the truncated range of ages observable from 1988 through 2005. These differences will tend to be smaller the narrower the age window considered.

For example, in the Figure below, we show two distributions of age at death in which population A has life expectancy at age 65 of 18 and population B has life expectancy at age 65 of 19.

If we only observe deaths from ages 78 to 95, as we would for the cohort of 1910, the difference in these truncated means will only be . . . , understating the $e(65)$ difference by a factor of X.

It is possible to estimate more sophisticated models that take into account truncation and provide parametric and other model-based estimates of the untruncated mortality distribution (Alexander 2018). This is particularly useful for estimating changes in differences over time, when the researcher does not want to confound time trends in the effects of covariates with changing ages of truncation.

The regression approach has the advantage of being simpler, faster and still easy to interpret. In order to translate regression results, we recommend using a multipicative adjustment factor, estimated using simulation.

The simulation assumes two Gompertz mortality schedules with the same senescence parameter b but differing modal ages of death M, such that their life expectancy at age 65 differs by 1 year. A function in the computer language `R` , shown below, produces estimates of the adjustment factor needed to translate differences in truncated means to differences in $e(65)$.

```
get.bunmd.adjust.factor(byear.vec,
                        b = 1/10,
                        M = 84,
                        e65.diff = 1,
                        N = 1 * 10^6)
```

The function allows the user to specify the set of birth cohorts, e.g. `byear.vec = 1895:1920`. It also allows the user to modify the baseline mortality schedule paramters b and M, as well as the simulated difference in $e(65)$. We find that the estimates are not sensitive to the choice of difference in $e(65)$, an encouraging result that permits use of the same adjustment factor to a range of observed differences in truncated means.

The table below shows the adjustment factors implied by different choices of b and M. The Human Mortality Database tells us that the birth cohort of 1910 had a modal age of death of about 84 and a senescence rate β of about 0.1. So, we recommend these as default choices. As can be seen from the table, the adjustment factors are typically about 3, with much larger or smaller values found only in the extreme cases. When the distribution of mortality is very compressed (high b) and the modal age is young, the truncated tails are smaller and the adjustment factor is closer to 2. When the distribution of mortality is disperse and the modal age is high, there is a lot of truncation, and the adjustment factor can be 5 or even higher.

Regression coefficient adjustment factors for combined birth cohorts 1895 to 1920, estimated by simulation for different Gompertz parameter values b and M.

		b				
		0.08	0.09	0.1	0.11	0.12
M	77	3.0	2.5	2.2	2.1	1.9
	78	3.2	2.7	2.4	2.2	2.0
	79	3.3	2.9	2.5	2.4	2.1
	80	3.5	3.0	2.8	2.4	2.2
	81	3.9	3.2	2.8	2.5	2.4
	82	4.1	3.3	3.1	2.6	2.5
	83	4.3	3.7	3.2	2.9	2.6
	84	4.7	3.8	3.4	3.1	2.7
	85	4.9	3.8	3.6	3.2	2.9
	86	5.6	4.6	3.9	3.4	3.0
	87	6.1	5	4.2	3.7	3.3

We recommend that users estimate the appropriate adjustment factor for their analysis by specifying their choice of birth cohorts. The Gompertz parameters can be left at their default values of $\beta = 0.1$ and $M = 84$ unless there is a reason to override these values based on external estimates.

Users can do their full statistical analysis, including hypothesis testing and model selection, using regression on age at death with dummy variables for year of birth. The adjustment factors can then be applied when discussing the magnitude of results and comparing them to other research findings. When greater precision is desired, or comparisons among cohorts are made, then more complex methods are needed.

Geography

There are several geography variables in the BUNMD. The Social Security application entries include information on birthplace. For persons born in the United States, the geographic resolution is state-level, and for persons born outside of the United States, the geographic resolution is country-level. The Numident death entries contains the 9-digit ZIP Code of residence at the time of death for a portion of records. ZIP Codes, while not the most robust geographic unit of analysis, can offer insights into a variety of spatial questions (Grubesic and Matisziw 2006).

Figure 5 shows life expectancy at age 65 for the birth cohort of 1900 in Ohio's Cuyahoga County by ZIP Code. Life expectancy is lower in inner-city Cleveland, and higher in its surrounding suburbs. These old-age mortality disparities are likely driven by racial segregation.

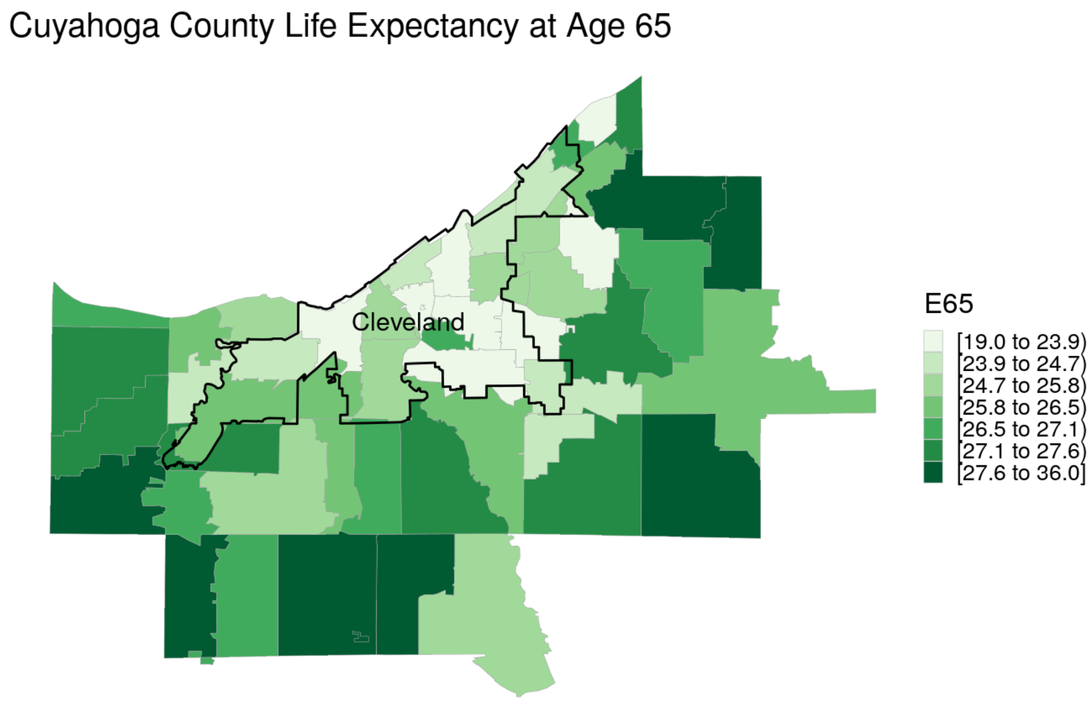

Figure 5: Life expectancy at age 65 in Cuyahoga County for the birth cohort of 1900.

Gompertz maximum likelihood estimation: race differentials

Gompertz maximum likelihood estimation can also be used to look at race differentials in mortality by state in the BUNMD. This method combines the observed distribution of deaths over a certain window of deaths with external knowledge of human mortality age-patterns, allowing us to estimate mortality rates given a truncated window of deaths.

We are assuming that the Gompertz model is appropriate and that the deaths we observe reflects the true population cohort distribution. Under-c verage will not bias estimates as long as the under-coverage is happening at random.

In Figure 7, we compare estimates of life expectancy at age 65 for Whites and Blacks over time for the cohorts of 1900 to 1920 in the state of Alabama using a Gompertz model. The size of the BUNMD allows researchers to identify heterogeneity and identify patterns of mortality obscured by composite population patterns (Vaupel and Yashin 1985).

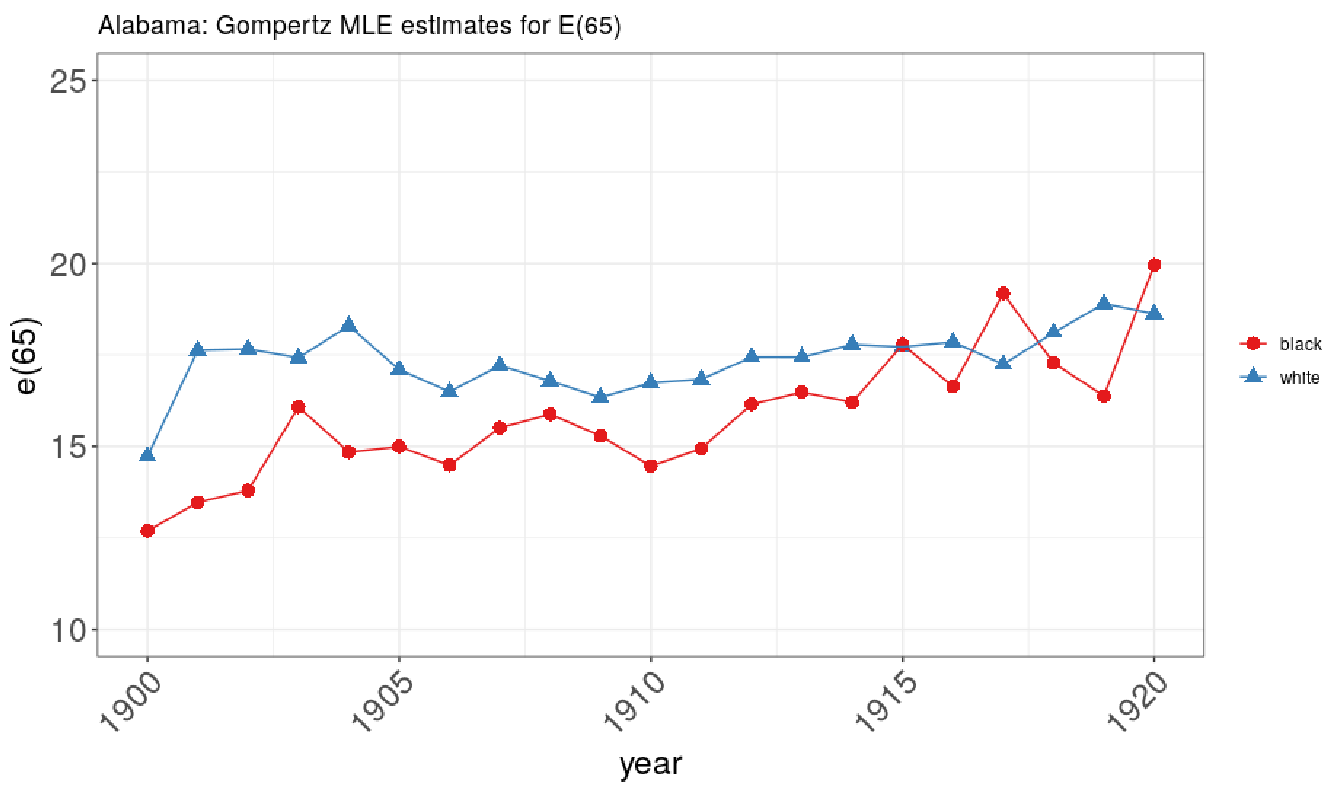

Figure 6: Gompertz E65 estimates for Alabama for Whites and Blacks.

Conclusion

The NARA Numident release has created a new administrative data resource for researchers studying mortality. We introduce the BUNMD, a cleaned and harmonized version of the NARA Numident records with over 49-million deaths. We provide an overview of statistical methods for estimating mortality using this deaths-only dataset. The high spatial resolution and demographic covariates open up new avenues for high-resolution mortality research, and the open-access nature of the data ensures that research is reproducible and extendable.

Public distribution, acknowledgement, conditions

The authors benefited from helpful discussions with Lynn Goodsell, guy from SSA, Berkeley HMD, etc. TODO.

The original SS-5 Files are available for download at _______________?

References

47, Record Group. 2019. "Numerical Identification (NUMIDENT) Files Frequently Asked Questions." National Archives; Records Administration.

Alexander, Monica. 2018. "Deaths Without Denominators: Using a Matched Dataset to Study Mortality Patterns in the United States." Preprint. SocArXiv. https://doi.org/10.31235/osf.io/q79ye.

Black, Dan, (first), Hsu Yu-Chieh, and Lynne Steuerle Schofield. 2001. "The Methuselah Effect: The Pernicious Impact of Unreported Deaths on Old Age Mortality Estimates," 45.

Chetty, Raj, Michael Stepner, Sarah Abraham, Shelby Lin, Benjamin Scuderi, Nicholas Turner, Augustin Bergeron, and David Cutler. 2016. "The Association Between Income and Life Expectancy in the United States, 2001-2014." *JAMA* 315 (16): 1750. https://doi.org/10.1001/jama.2016.4226.

Gavrilov, Leonid A, and Natalia S Gavrilova. 2012. "Mortality Measurement at Advanced Ages: A Study of the Social Security Administration Death Master File," 26.

Goldstein, Joshua .R, Monica Alexander, Casey Breen, Andrea Miranda González, and Felipe Menares. 2019. "CenSoc Mortality File: Version 2.0." Berkeley: University of California.

Goldstein, Joshua R., and Guy Stecklov. 2016. "From Patrick to John F.: Ethnic Names and Occupational Success in the Last Era of Mass Migration." *American Sociological Review* 81 (1): 85–106. https://doi.org/10.1177/0003122415621910.

Greene, William H. 2005. "Censored Data and Truncated Distributions." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.825845.

Grubesic, Tony H, and Timothy C Matisziw. 2006. "On the Use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data." *International Journal of Health Geographics* 5 (December): 58. https://doi.org/10.1186/1476-072X-5-58.

Harper, Sam, Richard F. MacLehose, and Jay S. Kaufman. 2014. "Trends in the Black-White Life Expectancy Gap Among US States, 1990–2009." *Health Affairs* 33 (8): 1375–82. https://doi.org/10.1377/hlthaff.2013.1273.

Hill, Mark E, and Ira Rosenwaike. 2001. "The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages." *Social Security Bulletin* 64 (1): 7.

Mehta, Neil K., Irma T. Elo, Michal Engelman, Diane S. Lauderdale, and Bert M. Kestenbaum. 2016. "Life Expectancy Among U.S.-born and Foreign-Born Older Adults in the United

States: Estimates from Linked Social Security and Medicare Data." *Demography* 53 (4): 1109–34. https://doi.org/10.1007/s13524-016-0488-4.

Newby, C. Alison, and Julie A. Dowling. 2007. "Black and Hispanic: The Racial Identification of Afro-Cuban Immigrants in the Southwest." *Sociological Perspectives* 50 (3): 343–66. https://doi.org/10.1525/sop.2007.50.3.343.

Puckett, Carolyn. 2009. "The Story of the Social Security Number." *Social Security Bulletin* 69 (2): 21.

Ruggles, Steven. 2014. "Big Microdata for Population Research." *Demography* 51 (1): 287–97. https://doi.org/10.1007/s13524-013-0240-2.

Schöley, Jonas, and Frans Willekens. 2017. "Visualizing Compositional Data on the Lexis Surface." *Demographic Research* 36 (February): 627–58. https://doi.org/10.4054/DemRes.2017.36.21.

Scott, Charles G. 1999. "Identifying the Race or Ethnicity of SSI Recipients," 12.

Vaupel, James W, and Anatoli I Yashin. 1985. "Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics." *The American Statisticiam* 39: 11.

Wachter, Kenneth. 2014. *Essential Demographic Methods*. Harvard University Press.

Waldron, Hilary. 2007. "Trends in Mortality Differentials and Life Expectancy for Male Social Security-Covered Workers, by Socioeconomic Status." *Social Security Bulletin* 67 (3): 28.