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

Joshua Goldstein*
Casey Breen†

22 May, 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 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.

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

# Contents

# 1. Introduction

The Numerical Identification System (Numident) forms the backbone of the U.S. Social Security Administration's recordkeeping 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 release of 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 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 an 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, birthplaces, 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. 2016). 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 subpopulations, and small geographic areas. The large sample sizes also enable the study of mortality at the oldest ages, when there are only a 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 on the use of this public data, researchers can also link records to their own restricted datasets.

The NARA Numident records pose a challenge for mortality estimation. Because the dataset 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.

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 complete count 1940 Census, to create a rich, publicly linked administrative dataset for the study of mortality (Goldstein et al. 2019).

## 2. The Structure and Content of the NARA Numident 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 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, corresponding to name changes each time she got married.

Table 1: Constructing the BUNMD from NARA Numident Records

## Thurgood Marshall

| | ssn | fname | lname | birth date | | | | | | | sex | race | | bpl | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Application Entry 1** | 131074264 | THURGOOD | MARSHALL | 7/2/1908 | | | | | | | 1 | **2** | | **MD** | | |
| | **ssn** | **fname** | **lname** | **birth date** | | | **death date** | | | | **sex** | | | | **zip_residence** | |
| **Death Entry** | **131074264** | **THURGOOD** | **MARSHALL** | **7/2/1908** | | | **1/24/1993** | | | | **1** | | | | **220411335** | |
| | **ssn** | **fname** | **lname** | **byear** | **bmonth** | **bday** | **dyear** | **dmonth** | **dday** | **death_age** | **sex** | **race_first** | **race_last** | **bpl** | **zip_residenc** | **number_apps** |
| **BUNMD Entry** | **131074264** | **THURGOOD** | **MARSHALL** | **1908** | **7** | **2** | **1993** | **1** | **24** | **84*** | **1** | **2** | **2** | **2400** | **220411335** | **1*** |

## Lana Turner

| | ssn | fname | lname | birth date | | | | | race | | sex | | | bpl | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Application Entry 1** | 567183907 | LANA | TURNER | 2/8/1921 | | | | | **1** | | 2 | | | ID | | |
| **Application Entry 2** | 567183907 | LANA | TOPPING | 2/8/1921 | | | | | 1 | | 2 | | | **ID** | | |
| **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** | | | | **zip_residence** | |
| **Death Entry** | **567183907** | **LANA** | **TURNER** | **2/8/1921** | | | **6/29/1995** | | | | **2** | | | | **900255240** | |
| | **ssn** | **fname** | **lname** | **byear** | **bmonth** | **bday** | **dyear** | **dmonth** | **dday** | **death_age** | **sex** | **race_first** | **race_last** | **bpl** | **zip_residenc** | **number_apps** |
| **BUNMD Entry** | **567183907** | **LANA** | **TURNER** | **1921** | **2** | **8** | **1995** | **6** | **29** | **74*** | **2** | **1** | **1** | **1600** | **900255240** | **4*** |

*Note*: Bolded values were selected for in the BUNMD. Starred values represent contructed variables not in the original records. Various features of the BUNMD creation algorithm can been seen here. For example, we select a person's first and last name from their death entries. We select the race and birthplace (bpl) from the application records. We use a crosswalk to recode the original two-letter character birthplace codes into a numeric code schema. We select race information from the application files to construct the race_first and race_last variables. The death_age and number_apps variables are not included in the original records but were constructed post-hoc using information in the original records.

The NARA Numident records contain three types of entries: application, claim, and death. The application entries contained information extracted from one of two forms, either the "Application for a Social Security Card" or the "Application for Social Security Account Number." The application entries contain a person's full name, race, sex, birthplace, date of birth, parents' full names, and other administrative information. Individuals may submit additional applications to replace a lost Social Security card, fix an error in a previous entry, or make a name change. Claim entries contain a person's full name, date of birth, sex, and whether the claim was a life claim or a death claim. Some individuals may have multiple claim entries. 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.

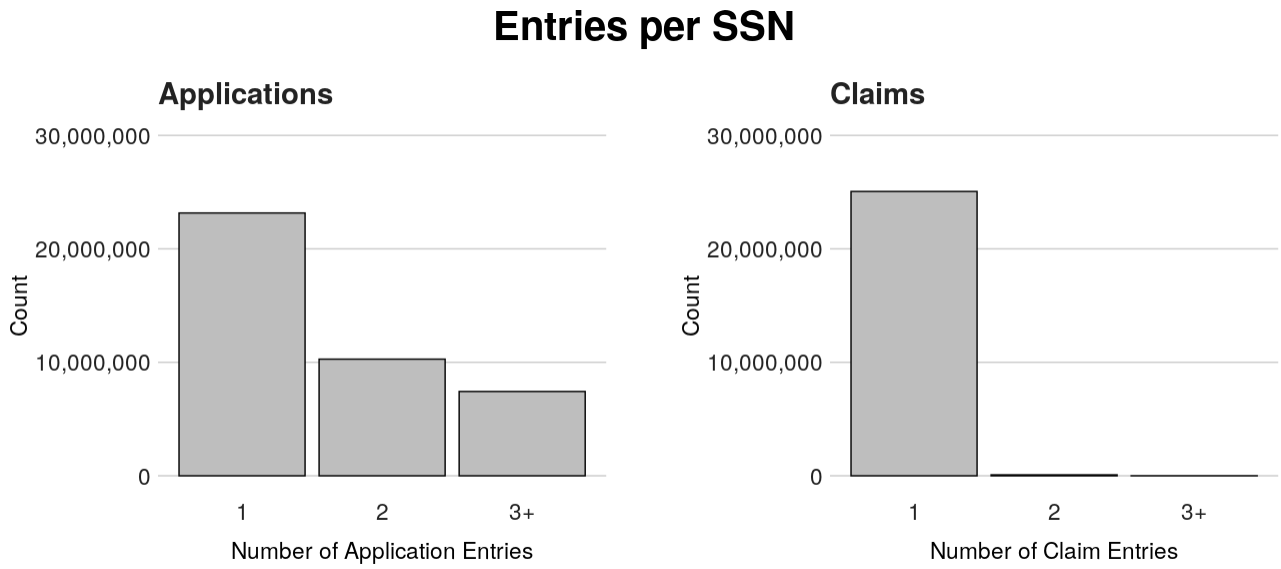



Figure 1: Number of entries per person for the Numident Application and Claim files.

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 claim records into a single file with one record per person. This file is available for download at: https://censoc-download.demog.berkeley.edu/. The file includes about 49 million records and 28 variables. It is 5.7GB in size.

The original NARA release contained 49,459,293 death record entries, 72,120,516 application entries for 40,870,455 unique persons, and 25,228,257 claim 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 claim entries. For individuals with multiple application or claim 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, and (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.

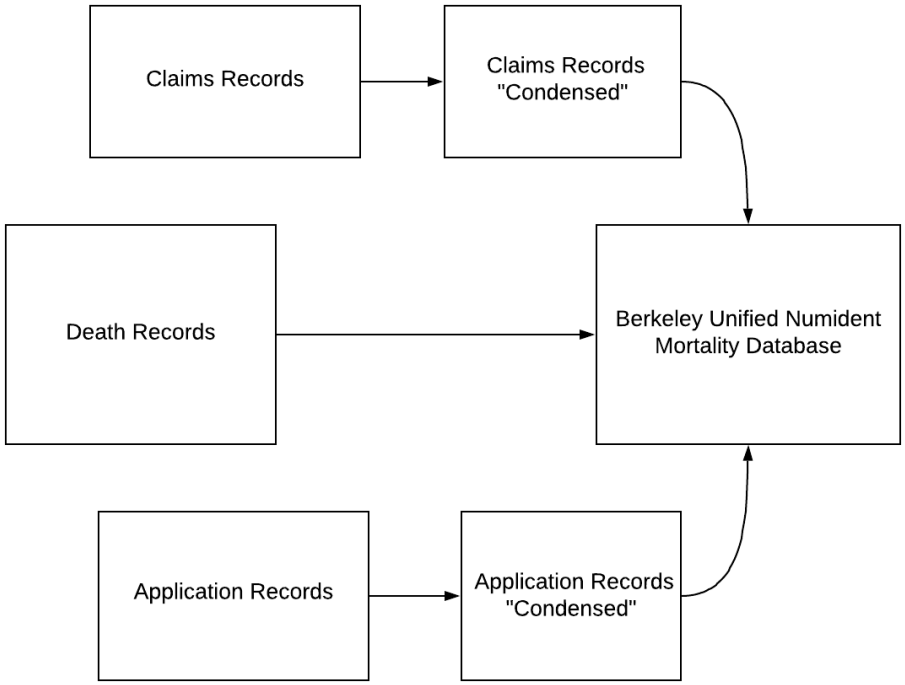



Figure 2: Berkeley Unified Numident Mortality Database creation flowchart.

### 2.1 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 process used by the Social Security Administration 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" (Record Group 47 2019). However, there are many individuals who fit those criteria who are not included in the dataset. Figure 3 compares the

total number of deaths for persons age 65+ (when coverage is highest) 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 — the "complete case" sample — 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.

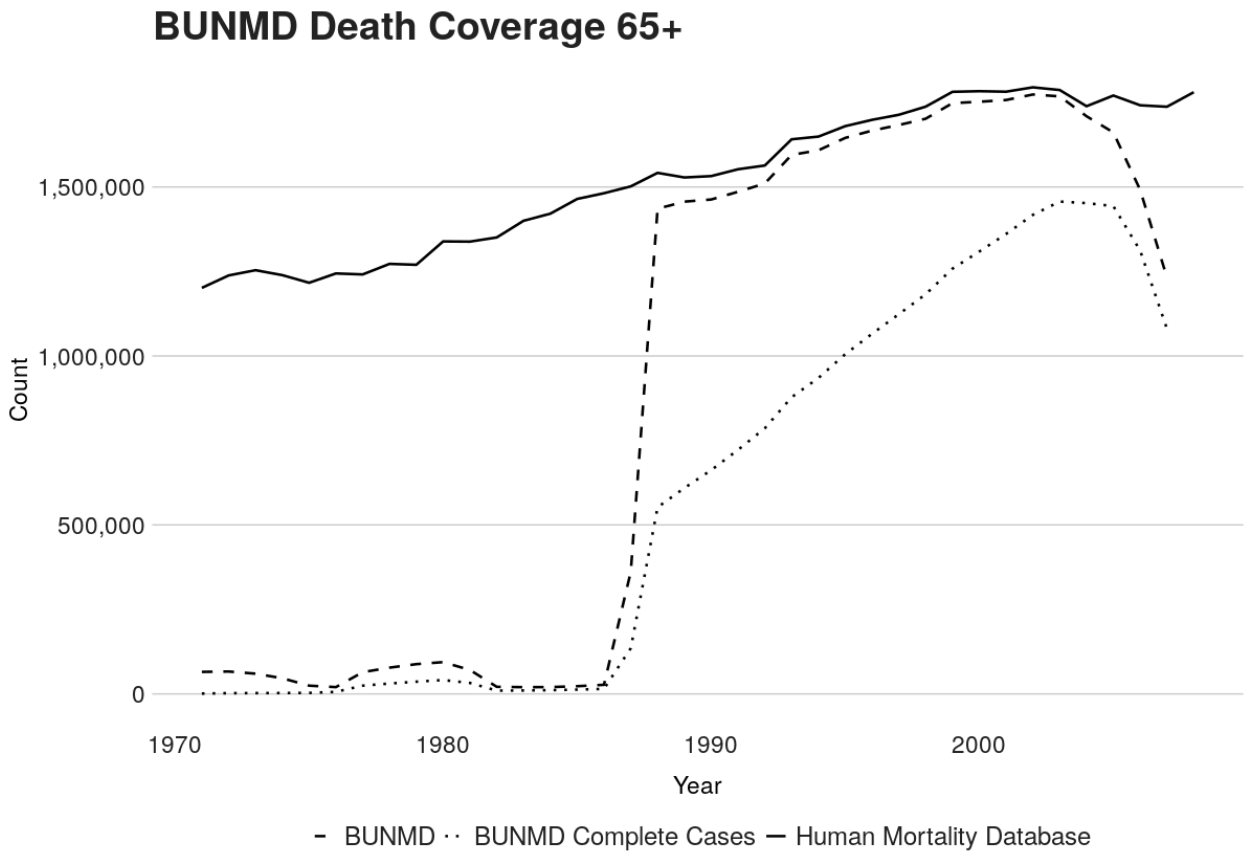



Figure 3: BUNMD Death Coverage for persons 65+. Coverage is most complete in the BUNMD for the window of 1988 to 2005.

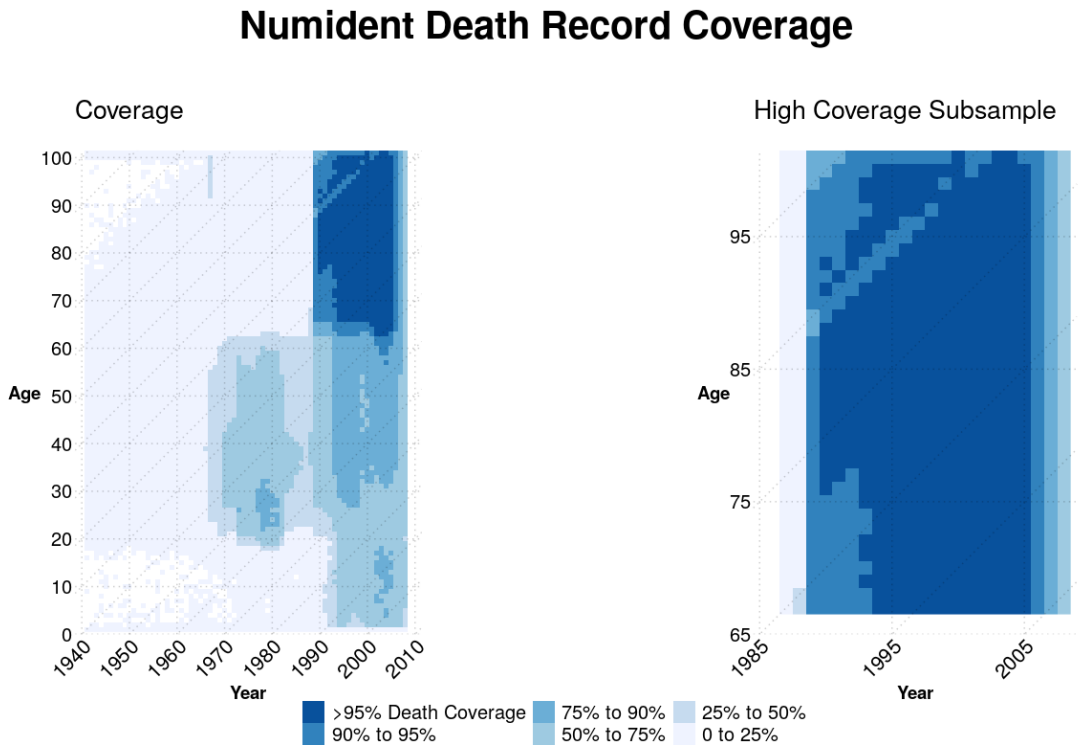



Figure 4: Lexis diagrams of BUNMD death coverage. The left panel shows the BUNMD death coverage for 1940 to 2010. The right panel zooms in on age-period area where coverage is highest.

## 3. 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.

### 3.1 Method 1: Parametric Survival Models (for Truncated Data)

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

$$h(x) = a \cdot e^{bx} \tag{1}$$

The constant exponential rate of increase is most pronounced for the ages of 70 to 90. 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:

$$h(x) = c + a \cdot e^{(bx)} \tag{2}$$

Although there is still much debate, at older ages 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 $a$ and known right truncation $b$, we can define the conditional distribution to be

$$f_{trunc} = \frac{f_\theta(x)}{\int_a^b f_\theta(x)dx} = \frac{f_\theta(x)}{F_\theta(b) - F_\theta(a)} \tag{3}$$

with likelihood of

$$L(\theta|X) = \prod \frac{f_\theta(x_i)}{F_\theta(b) - F_\theta(a)} \tag{4}$$

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

### 3.2 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:

$$\text{Age\_at\_death} = \text{birth\_year\_dummy} + \text{covariates of interest} \tag{5}$$

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 downwardly bias the estimated effects of any covariates (Greene 2005). Truncation excludes the tails of the distribution, thus reducing the average difference between groups; 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. Figure 5 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, as in the example comparing men and women, found using the data from the cohort of 1910 (observed from 1988 to 2005) translates to an $e65$ difference of 0.5 x 2.3 = 1.15 years.

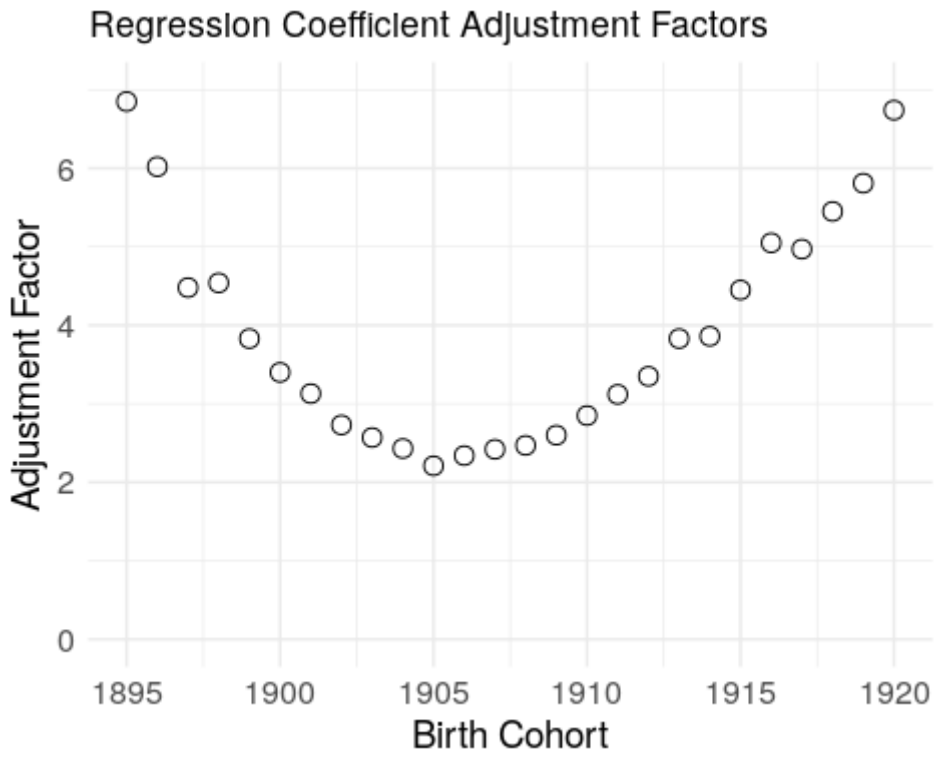



Figure 5: Regression coefficient adjustments for birth cohorts of 1895-1920 using baseline mortality schedule parameters $b = 1/10$ and $M = 84$

Table 3 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 $\beta$ of about 0.1. 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 dispersed and the modal age is high, there is significant truncation, and the adjustment factor can be 5 or even higher.

| | | $b$ | | | | |
|---|---|---|---|---|---|---|
| | | 0.08 | 0.09 | 0.1 | 0.11 | 0.12 |
| | 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 |
| $M$ | 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 |

Table 3: Regression coefficient adjustment factors for combined birth cohorts 1895 to 1920, estimated by simulation for different Gompertz parameter values $b$ and $M$.

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.

We show regression results that 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 6 below, we show two distributions of age at death in which population A has an life expectancy at age 65 ($e65$) of 18 years and population B has an $e65$ of 19 years.


Figure 6: Distributions of age of death for two populations. Gray lines represent hypothetical left and right truncation.

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 0.358, understating the $e65$ difference by a factor of 2.79.

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 multiplicative 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 $e65$ differs by one 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 $e65$.

```
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 parameters $b$ and $M$, as well as the simulated difference in $e65$. We find that the estimates are not sensitive to the choice of difference in $e65$, an encouraging result that permits use of the same adjustment factor to a range of observed differences in truncated means.

## 4. Case Studies

### 4.1 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 ZIP Code 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 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."

### 4.2 Country-of-Birth Differences

To measure mortality differences by country of birth, we have chosen the 19 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 7 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 $e65$. A good approximation for translating the differences in the truncated sample $e65$ is to multiply the regression coefficients by about 4. We discuss how such multiplicative factors can be estimated in the section on Ordinary Least Squares Regression.

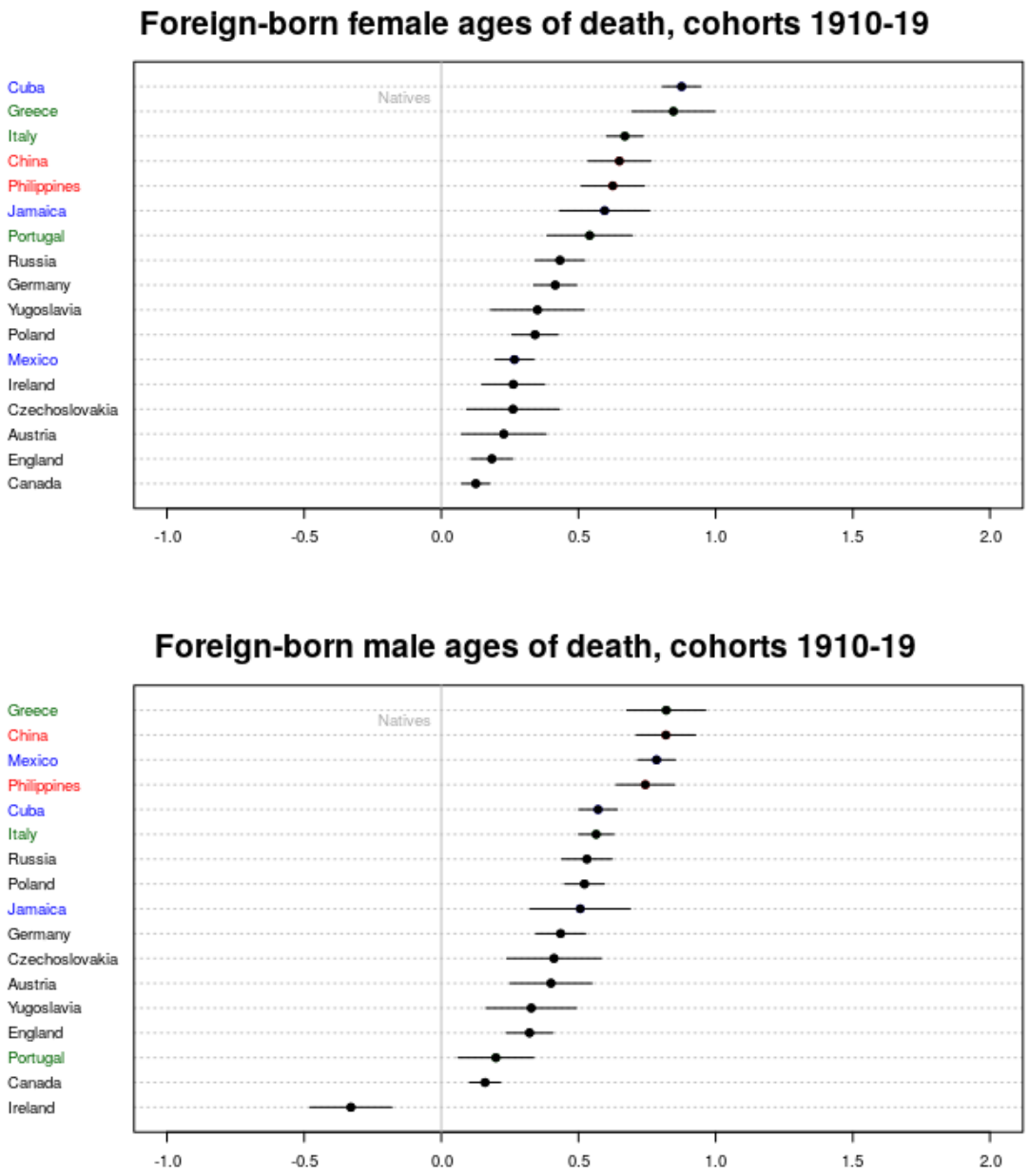



Figure 7: Difference in mean age of death between natives and foreign-born.

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, the 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 explored 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).

### 4.3 Racial Differences Among Cuban Immigrants

We saw above that Cuban immigrants are among the longest-living subgroups in the United States. We can use the NARA Numident to further explore whether immigrants from this racially diverse country differ in their longevity in the United States, keeping in mind that racial identity is self-identified 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 Table 3 below. We can see that Black Cuban immigrants died earlier than White Cuban immigrants. The effect of -0.7 years in the regression corresponds to an $e65$ of about 3 years. There is not a statistically significant 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 that are racially diverse. A deeper dive into the original SS-5 application files may also reveal interesting patterns about who chooses a Hispanic identity

and when. For example, one could link the Numident records to the 1940 Census, finding that those who change their identity from Hispanic to White tend to have higher earnings and educational attainment.

### 4.4 Geography

There are several geographic 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 the United States, the geographic resolution is country-level. The Numident death entry contains the 9-digit ZIP Code of residence at the time of death for some of the records. ZIP Codes, while not necessarily the most robust geographic unit of analysis, can offer insights into a variety of spatial questions (Grubesic and Matisziw 2006).

Figure 8 shows $e65$ for the birth cohorts of 1910-1919 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.

Table 1

| | *Dependent variable:* |
|---|---|
| | death_age |
| Race: Black | $-0.696^{***}$ (0.159) |
| Race: Other | 0.284 (0.159) |
| Byear1895 | $7.354^{**}$ (2.562) |
| Byear1897 | 4.835 (3.574) |
| Byear1898 | $8.932^{***}$ (1.194) |
| Byear1899 | 3.816 (2.027) |
| Byear1900 | $6.660^{***}$ (0.885) |
| Byear1901 | $5.082^{***}$ (1.010) |
| Byear1902 | $4.273^{***}$ (0.542) |
| Byear1903 | $4.597^{***}$ (0.394) |
| Byear1904 | $4.661^{***}$ (0.301) |
| Byear1905 | $3.213^{***}$ (0.253) |
| Byear1906 | $2.522^{***}$ (0.194) |
| Byear1907 | $2.311^{***}$ (0.174) |
| Byear1908 | $1.352^{***}$ (0.144) |
| Byear1909 | 0.206 (0.122) |
| Byear1911 | $-0.425^{***}$ (0.106) |
| Byear1912 | $-1.125^{***}$ (0.112) |
| Byear1913 | $-1.873^{***}$ (0.113) |
| Byear1914 | $-2.581^{***}$ (0.118) |
| Byear1915 | $-3.444^{***}$ (0.126) |
| Byear1916 | $-4.211^{***}$ (0.129) |
| Byear1917 | $-5.105^{***}$ (0.129) |
| Byear1918 | $-5.825^{***}$ (0.127) |
| Byear1919 | $-6.728^{***}$ (0.128) |
| Byear1920 | $-7.728^{***}$ (0.128) |
| Gender (Female) | $1.743^{***}$ (0.049) |
| Constant | $84.422^{***}$ (0.082) |
| Observations | 39,852 |
| $R^2$ | 0.280 |
| Adjusted $R^2$ | 0.280 |
| Residual Std. Error | 6.708 (df = 39824) |
| F Statistic | $574.450^{***}$ (df = 27; 39824) |
| *Note:* | $^{*}$p<0.05; $^{**}$p<0.01; $^{***}$p<0.001 |

Table 3: Regression results for mean age at death of Cuban immigrants by race, birth year, and sex. Reference group for this regression is White men born in 1910.

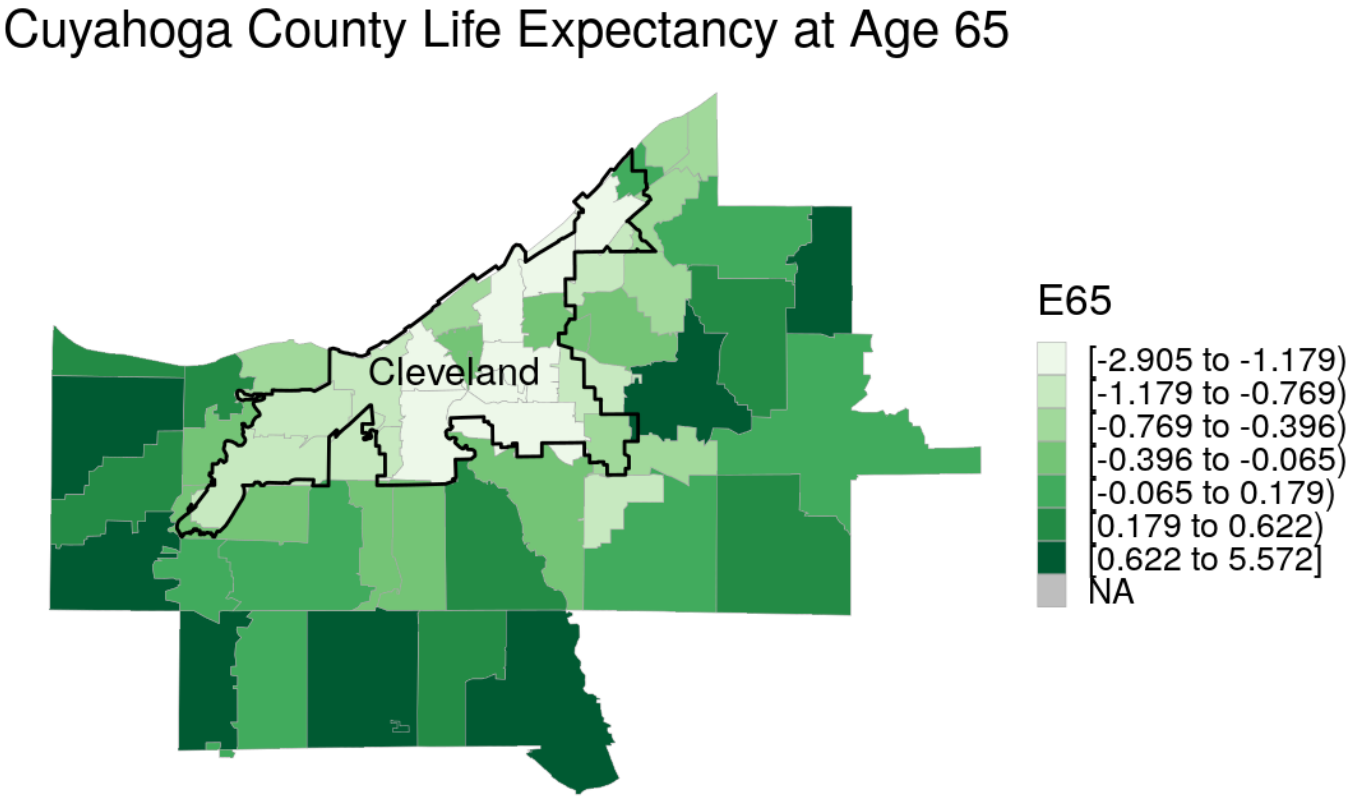



Figure 8: Difference in $e65$ in Cuyahoga County for the birth cohorts of 1910-1919

### 4.5 Gompertz Maximum Likelihood Estimation: Race Differentials

Gompertz maximum likelihood estimation (MLE) 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 (Alexander 2018). We are assuming that the Gompertz model is appropriate and that the deaths we observe reflect the true population cohort distribution. Undercoverage will not bias estimates as long as the undercoverage is happening at random.

In Figure 9, we compare estimates of $e65$ 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).

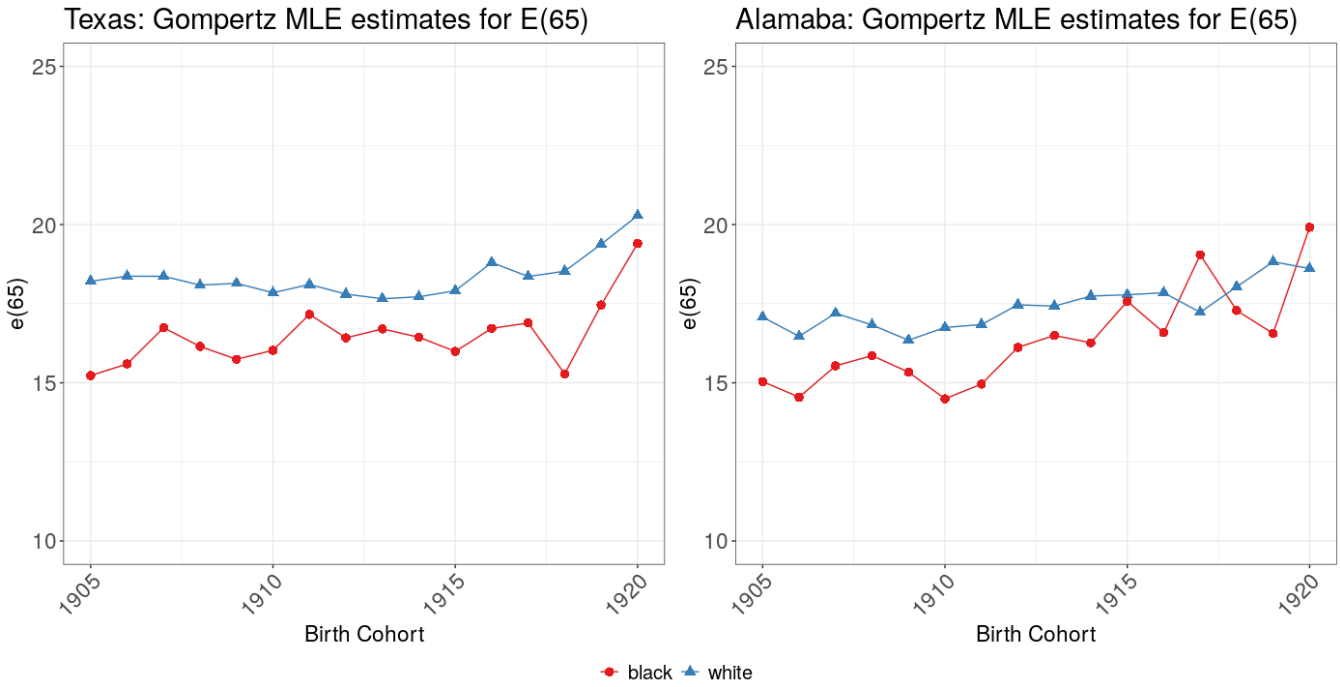



Figure 9: Gompertz $e65$ estimates for Alabama for Whites and Blacks.

We chose Alabama and Texas as they had large enough subpopulations, after disaggregation by birth cohort, for the Gompertz MLE approach to produce reasonable estimates. Developing diagnostics to assess the fit of the Gompertz MLE approach and estimate uncertainty is an area for future research.

### 4.7 Flu Seasons

The BUNMD's individual death counts by day allows researchers to study who was hit hardest by the flu, by ZIP Code, race, exact date of birth, and more. Figure 10 shows the four US big flu seasons at the end of the 1990s.

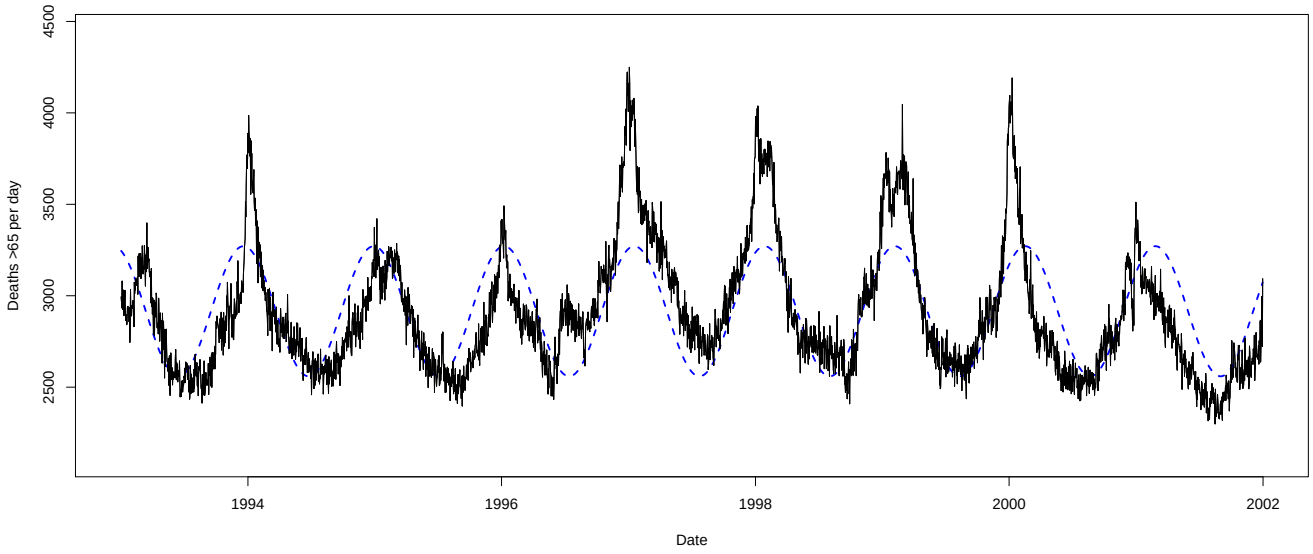



Figure 10: Death counts per day, 1993-2002.

## 5. Conclusion

The National Archives release of Numident records has created a new administrative data resource for researchers studying mortality. We have created a single-file, cleaned and harmonized version of this data, the BUNMD, covering more than 49 million deaths. In this paper we have described the original data and the steps we took to create BUNMD. We have also provided an overview of statistical methods for estimating mortality using this deaths-only data set and provided a series of examples showing how to use the BUNMD. Our hope is that these public administrative records, with their extraordinary spatial resolution and size, will open up new avenues for high-resolution mortality research. Furthermore, we hope that the open-access nature of the data will help foster research that is reproducible and extendable.

### Public distribution, acknowledgement, conditions

The authors benefited from helpful discussions with Lynn Goodsell, Berkeley Human Mortality Database, and Ugur Yildirim.

The original NARA Numident Records are available upon request.

## References


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.

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

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.

# Technical Appendix

This supplementary appendix presents the procedure for constructing the Berkeley Unified Numident Mortality Database (BUNMD) from the original NARA Numident records. Please see the BUNMD Codebook for variable descriptions, value labels, and tabulations.

## 1.1 NARA Numident Records

In 2013, the Social Security Administration transferred a set of Numident records to the National Archives (NARA). In 2019, we obtained the NARA Numident records and their accompanying documentation. The NARA Numident records are a subset of the records in the complete Numident. The NARA Numident records contain three types of entries: application, claim, and death. NARA delivered each set of entries separately as a set 20 fixed-width .txt files (3 x 20 = 60 files in total).

## 1.2 NARA Numident Metadata

We obtained three documents from the National Archives Technical Documentation series:

(https://aad.archives.gov/aad/popup-tech-info.jsp?s=5057; accessed 11/28/2019):

- Application (SS-5) Records Layout
- Death Records Layout
- Claim Records Layout

The record layout documents contain variable descriptions, value labels, technical notes, and the start and end position for each variable in the 60 fixed-width .txt files.

## 1.3 Code

All data processing was was done in the `R` Statistical Programming Language. The code to construct the BUNMD from the NARA Numident records is available at "Github.com/caseybreen/wcensoc/".

## 1.4 Data Preprocessing

For each of the three entry types, we read in the 20 fixed-width .txt files using the column position specified in the record layout documents. We then appended the 20 files into a single file, creating a single file for each of the three entry types.

We took the following steps to clean each file:

1. We changed the variable names to be more concise and informative. For example, we renamed the "NUMI_SEX" variable to "sex".
2. We harmonized the different codes to represent a missing value ("Unknown", "Unk", "Un", and "0") to "NA."

## 1.5 Combining NARA Numident Records into the BUNMD

The goal in constructing the BUNMD was to combine the NARA Numident records into a single, harmonized file with one record per person. The original records contain over 100+ variables. Some are not of general interest to the research community, while others contain 99%+ missing values (as shown in Figures 2-4). We selected a set of general-interest variables with high completeness.

While a person can only have one death entry, they might have several application or claim entries; information may be reported several times. For example, sex is reported in the application, claim, and death entries. Occasionally, a person reports different values of sex, race, place of birth, etc. across entries. To handle this response inconsistency, we developed a set of decision rules to select a single value across entries (see Table 2). In order to study name changes, race changes, and other features, the original NARA Numident records are useful and are available upon request.

## 1.6 Geography

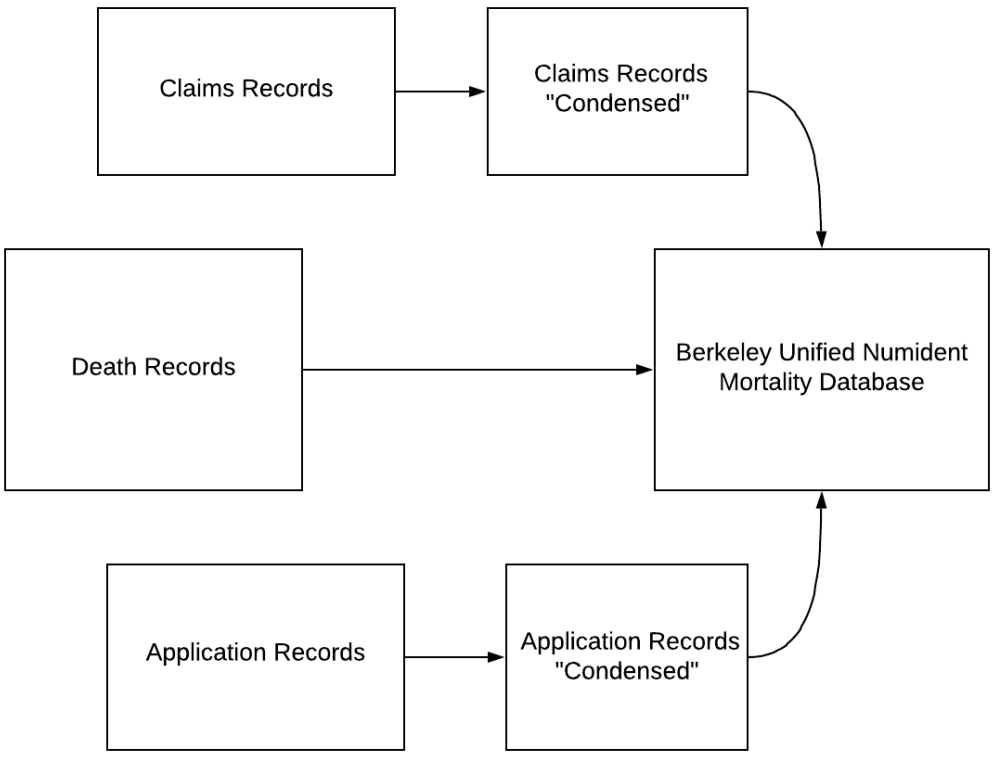
**Place of Birth:** There are several geographic variables in the NARA Numident records. The application entries have information on birthplace. For persons born in the United States, the geographic resolution is state-level. For persons born outside the United States, the geographic resolution is country-level. The NARA Numident uses two variables to convey birthplace. The first variable denotes whether a person was foreign born, and the second variable contains a two-letter state or country abbreviation. We harmonize these two

variables into one variable with a numeric coding schema. This coding schema matches the IPUMS-USA BPLD (Birthplace, detailed) schema.

**Place of Death:** The Numident death entry contains the 9-digit ZIP Code of the residence at the time of death. Sometimes, the full 9-digit ZIP Code is not available, and an "x" is used to represent a missing digit. This is the original convention used by the Social Security Administration.

**Social Security State:** The first three digits of a Social Security number correspond to the state in which a Social Security number was issued (prior to 1973) or to the ZIP Code of the mailing address listed in the Social Security application (after 1973). We constructed a variable "socstate" that reports the the state corresponding to the first three digits of the Social Security number. The Social Security Administration changed the assignment process in 2011 — after the last Social Security number for a person in the BUNMD was issued — and the first three digits no longer correspond to a state.

(https://www.ssa.gov/employer/stateweb.htm; accessed 11/28/2019)




(a)

Figure 11: BUNMD Construction Flow Chart

## 1.7 BUNMD Samples and Weights

We created two BUNMD samples with high death coverage. Sample 1 includes deaths to persons with high coverage (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 (HMD) on age at death, year of birth, year of death, and sex. We broke the sample into cells cross-classified by year of birth, year of death, age at death, and sex. We weighted each cell to the HMD "Deaths by Lexis triangles" totals. This allows aggregation to HMD totals by period or cohort.

$$W_j = \frac{\text{HMD deaths in cell j}}{\text{Numident Sample 1 deaths in cell j}} \tag{6}$$

## Decision Rules used for BUNMD

| Variable | Numident Source | Selection Rule |
|---|---|---|
| ssn | Death Entry | - |
| fname | Death Entry | - |
| mname | Death Entry | - |
| lname | Death Entry | - |
| byear | Death Entry | - |
| bmonth | Death Entry | - |
| bday | Death Entry | - |
| dyear | Death Entry | - |
| dmonth | Death Entry | - |
| dday | Death Entry | - |
| zip_residence | Death Entry | - |
| sex | Death, Application, or Claim Entry | Last Recorded Sex |
| race_first | Application Entry | First Recorded Race |
| race_last | Application Entry | Last Recorded Race |
| bpl | Application or Claim Entry | Last Recorded BPL |
| father_fname | Application or Claim Entry | Maximum Characters |
| father_mname | Application or Claim Entry | Maximum Characters |
| father_lname | Application or Claim Entry | Maximum Characters |
| mother_fname | Application or Claim Entry | Maximum Characters |
| mother_mname | Application or Claim Entry | Maximum Characters |
| mother_lname | Application or Claim Entry | Maximum Characters |
| race_change | Constructed | - |
| death_age | Constructed | - |
| socstate | Constructed | - |
| age_first_app | Constructed | - |
| number_apps | Constructed | - |
| number_claims | Constructed | - |
| weight | Constructed | - |
| ccweight | Constructed | - |

Table 1: The selection rules used to construct the BUNMD. For a given variable, we selected values from the death record, if available. If a value wasn't available in the death record, we chose a value from the application record using selection rules. If it was not available in either the death or application entry, we selected it from the claim record.

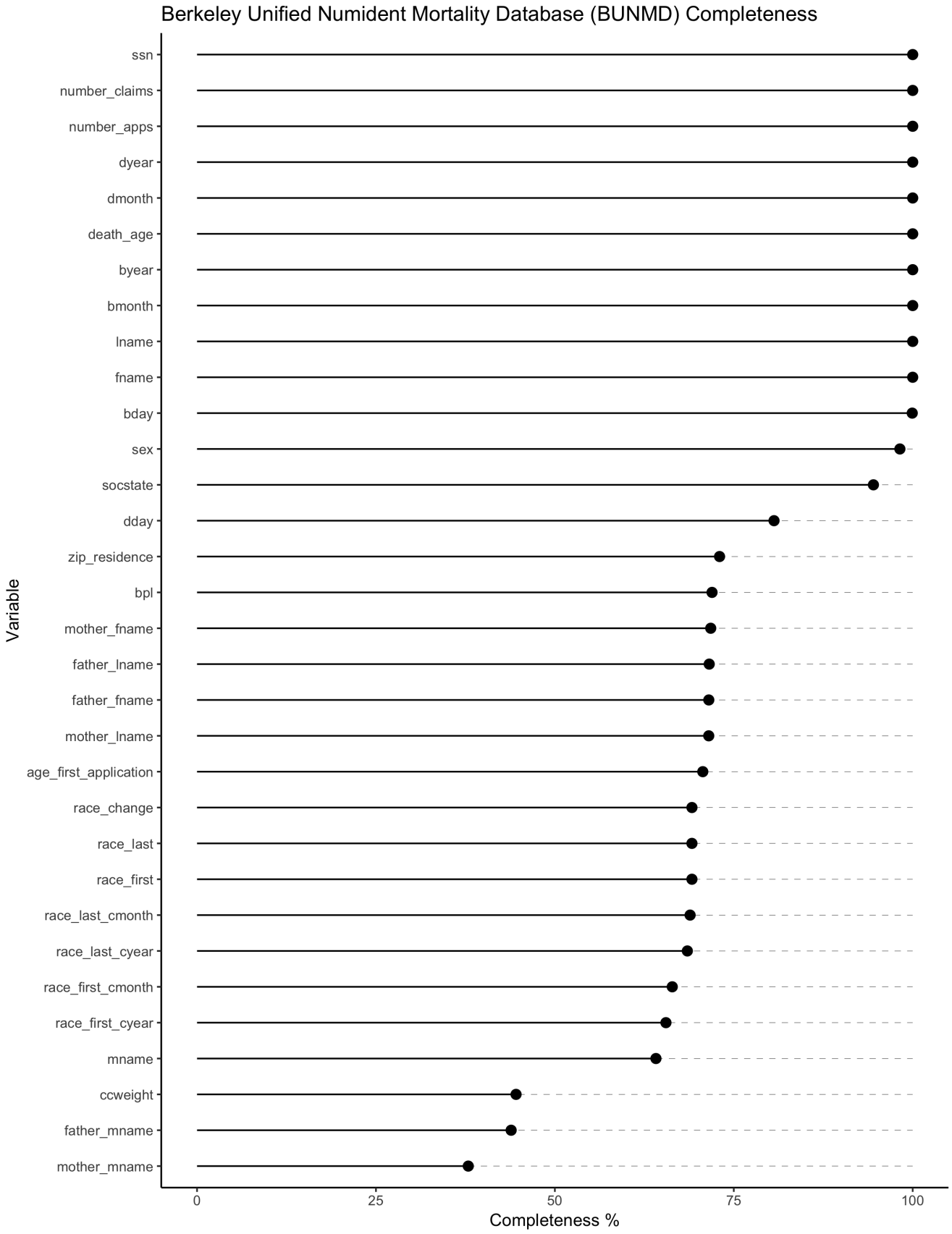

Berkeley Unified Numident Mortality Database (BUNMD) Completeness
Variable
ssn
number_claims
number_apps
dyear
dmonth
death_age
byear
bmonth
lname
fname
bday
sex
socstate
dday
zip_residence
bpl
mother_fname
father_lname
father_fname
mother_lname
age_first_application
race_change
race_last
race_first
race_last_cmonth
race_last_cyear
race_first_cmonth
race_first_cyear
mname
ccweight
father_mname
mother_mname
0
25
50
75
100
Completeness %

Figure 12

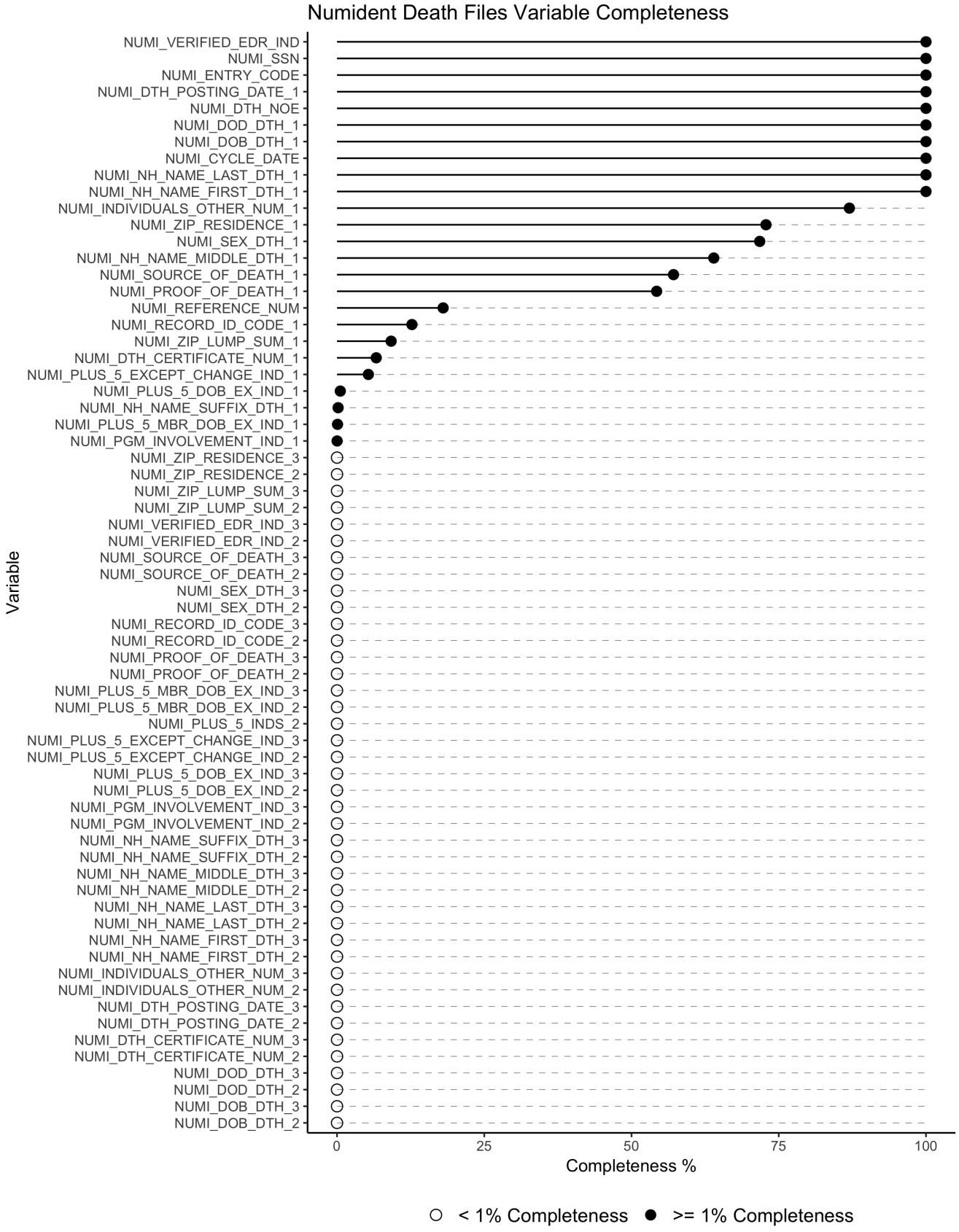

Numident Death Files Variable Completeness
NUMI_VERIFIED_EDR_IND
NUMI_SSN
NUMI_ENTRY_CODE
NUMI_DTH_POSTING_DATE_1
NUMI_DTH_NOE
NUMI_DOD_DTH_1
NUMI_DOB_DTH_1
NUMI_CYCLE_DATE
NUMI_NH_NAME_LAST_DTH_1
NUMI_NH_NAME_FIRST_DTH_1
NUMI_INDIVIDUALS_OTHER_NUM_1
NUMI_ZIP_RESIDENCE_1
NUMI_SEX_DTH_1
NUMI_NH_NAME_MIDDLE_DTH_1
NUMI_SOURCE_OF_DEATH_1
NUMI_PROOF_OF_DEATH_1
NUMI_REFERENCE_NUM
NUMI_RECORD_ID_CODE_1
NUMI_ZIP_LUMP_SUM_1
NUMI_DTH_CERTIFICATE_NUM_1
NUMI_PLUS_5_EXCEPT_CHANGE_IND_1
NUMI_PLUS_5_DOB_EX_IND_1
NUMI_NH_NAME_SUFFIX_DTH_1
NUMI_PLUS_5_MBR_DOB_EX_IND_1
NUMI_PGM_INVOLVEMENT_IND_1
NUMI_ZIP_RESIDENCE_3
NUMI_ZIP_RESIDENCE_2
NUMI_ZIP_LUMP_SUM_3
NUMI_ZIP_LUMP_SUM_2
NUMI_VERIFIED_EDR_IND_3
NUMI_VERIFIED_EDR_IND_2
NUMI_SOURCE_OF_DEATH_3
NUMI_SOURCE_OF_DEATH_2
NUMI_SEX_DTH_3
NUMI_SEX_DTH_2
NUMI_RECORD_ID_CODE_3
NUMI_RECORD_ID_CODE_2
NUMI_PROOF_OF_DEATH_3
NUMI_PROOF_OF_DEATH_2
NUMI_PLUS_5_MBR_DOB_EX_IND_3
NUMI_PLUS_5_MBR_DOB_EX_IND_2
NUMI_PLUS_5_INDS_2
NUMI_PLUS_5_EXCEPT_CHANGE_IND_3
NUMI_PLUS_5_EXCEPT_CHANGE_IND_2
NUMI_PLUS_5_DOB_EX_IND_3
NUMI_PLUS_5_DOB_EX_IND_2
NUMI_PGM_INVOLVEMENT_IND_3
NUMI_PGM_INVOLVEMENT_IND_2
NUMI_NH_NAME_SUFFIX_DTH_3
NUMI_NH_NAME_SUFFIX_DTH_2
NUMI_NH_NAME_MIDDLE_DTH_3
NUMI_NH_NAME_MIDDLE_DTH_2
NUMI_NH_NAME_LAST_DTH_3
NUMI_NH_NAME_LAST_DTH_2
NUMI_NH_NAME_FIRST_DTH_3
NUMI_NH_NAME_FIRST_DTH_2
NUMI_INDIVIDUALS_OTHER_NUM_3
NUMI_INDIVIDUALS_OTHER_NUM_2
NUMI_DTH_POSTING_DATE_3
NUMI_DTH_POSTING_DATE_2
NUMI_DTH_CERTIFICATE_NUM_3
NUMI_DTH_CERTIFICATE_NUM_2
NUMI_DOD_DTH_3
NUMI_DOD_DTH_2
NUMI_DOB_DTH_3
NUMI_DOB_DTH_2
Variable
0
25
50
75
100
Completeness %
< 1% Completeness
>= 1% Completeness

Figure 13

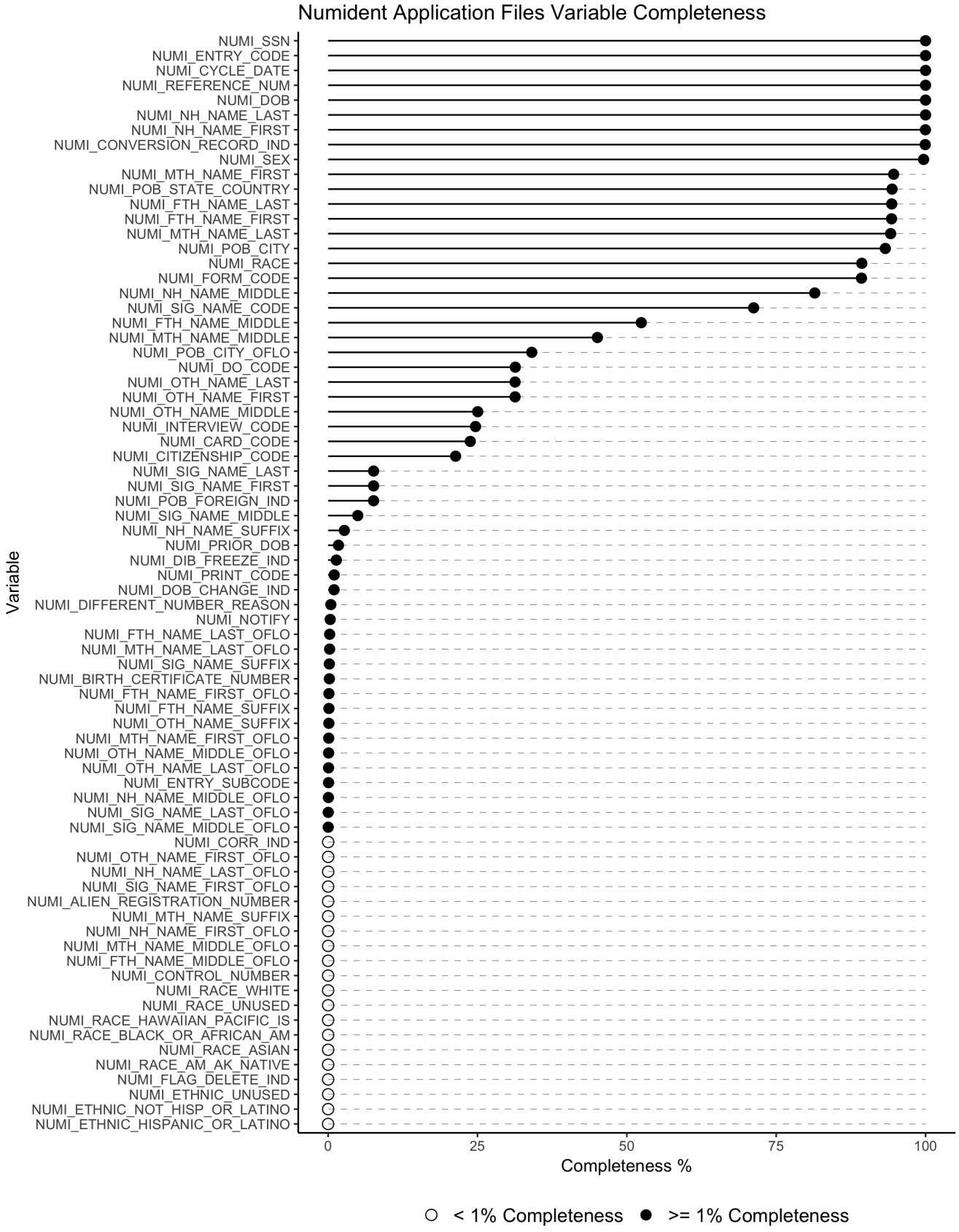

Numident Application Files Variable Completeness
Variable
NUMI_SSN
NUMI_ENTRY_CODE
NUMI_CYCLE_DATE
NUMI_REFERENCE_NUM
NUMI_DOB
NUMI_NH_NAME_LAST
NUMI_NH_NAME_FIRST
NUMI_CONVERSION_RECORD_IND
NUMI_SEX
NUMI_MTH_NAME_FIRST
NUMI_POB_STATE_COUNTRY
NUMI_FTH_NAME_LAST
NUMI_FTH_NAME_FIRST
NUMI_MTH_NAME_LAST
NUMI_POB_CITY
NUMI_RACE
NUMI_FORM_CODE
NUMI_NH_NAME_MIDDLE
NUMI_SIG_NAME_CODE
NUMI_FTH_NAME_MIDDLE
NUMI_MTH_NAME_MIDDLE
NUMI_POB_CITY_OFLO
NUMI_DO_CODE
NUMI_OTH_NAME_LAST
NUMI_OTH_NAME_FIRST
NUMI_OTH_NAME_MIDDLE
NUMI_INTERVIEW_CODE
NUMI_CARD_CODE
NUMI_CITIZENSHIP_CODE
NUMI_SIG_NAME_LAST
NUMI_SIG_NAME_FIRST
NUMI_POB_FOREIGN_IND
NUMI_SIG_NAME_MIDDLE
NUMI_NH_NAME_SUFFIX
NUMI_PRIOR_DOB
NUMI_DIB_FREEZE_IND
NUMI_PRINT_CODE
NUMI_DOB_CHANGE_IND
NUMI_DIFFERENT_NUMBER_REASON
NUMI_NOTIFY
NUMI_FTH_NAME_LAST_OFLO
NUMI_MTH_NAME_LAST_OFLO
NUMI_SIG_NAME_SUFFIX
NUMI_BIRTH_CERTIFICATE_NUMBER
NUMI_FTH_NAME_FIRST_OFLO
NUMI_FTH_NAME_SUFFIX
NUMI_OTH_NAME_SUFFIX
NUMI_MTH_NAME_FIRST_OFLO
NUMI_OTH_NAME_MIDDLE_OFLO
NUMI_OTH_NAME_LAST_OFLO
NUMI_ENTRY_SUBCODE
NUMI_NH_NAME_MIDDLE_OFLO
NUMI_SIG_NAME_LAST_OFLO
NUMI_SIG_NAME_MIDDLE_OFLO
NUMI_CORR_IND
NUMI_OTH_NAME_FIRST_OFLO
NUMI_NH_NAME_LAST_OFLO
NUMI_SIG_NAME_FIRST_OFLO
NUMI_ALIEN_REGISTRATION_NUMBER
NUMI_MTH_NAME_SUFFIX
NUMI_NH_NAME_FIRST_OFLO
NUMI_MTH_NAME_MIDDLE_OFLO
NUMI_FTH_NAME_MIDDLE_OFLO
NUMI_CONTROL_NUMBER
NUMI_RACE_WHITE
NUMI_RACE_UNUSED
NUMI_RACE_HAWAIIAN_PACIFIC_IS
NUMI_RACE_BLACK_OR_AFRICAN_AM
NUMI_RACE_ASIAN
NUMI_RACE_AM_AK_NATIVE
NUMI_FLAG_DELETE_IND
NUMI_ETHNIC_UNUSED
NUMI_ETHNIC_NOT_HISP_OR_LATINO
NUMI_ETHNIC_HISPANIC_OR_LATINO
0
25
50
75
100
Completeness %
< 1% Completeness
>= 1% Completeness

Figure 14

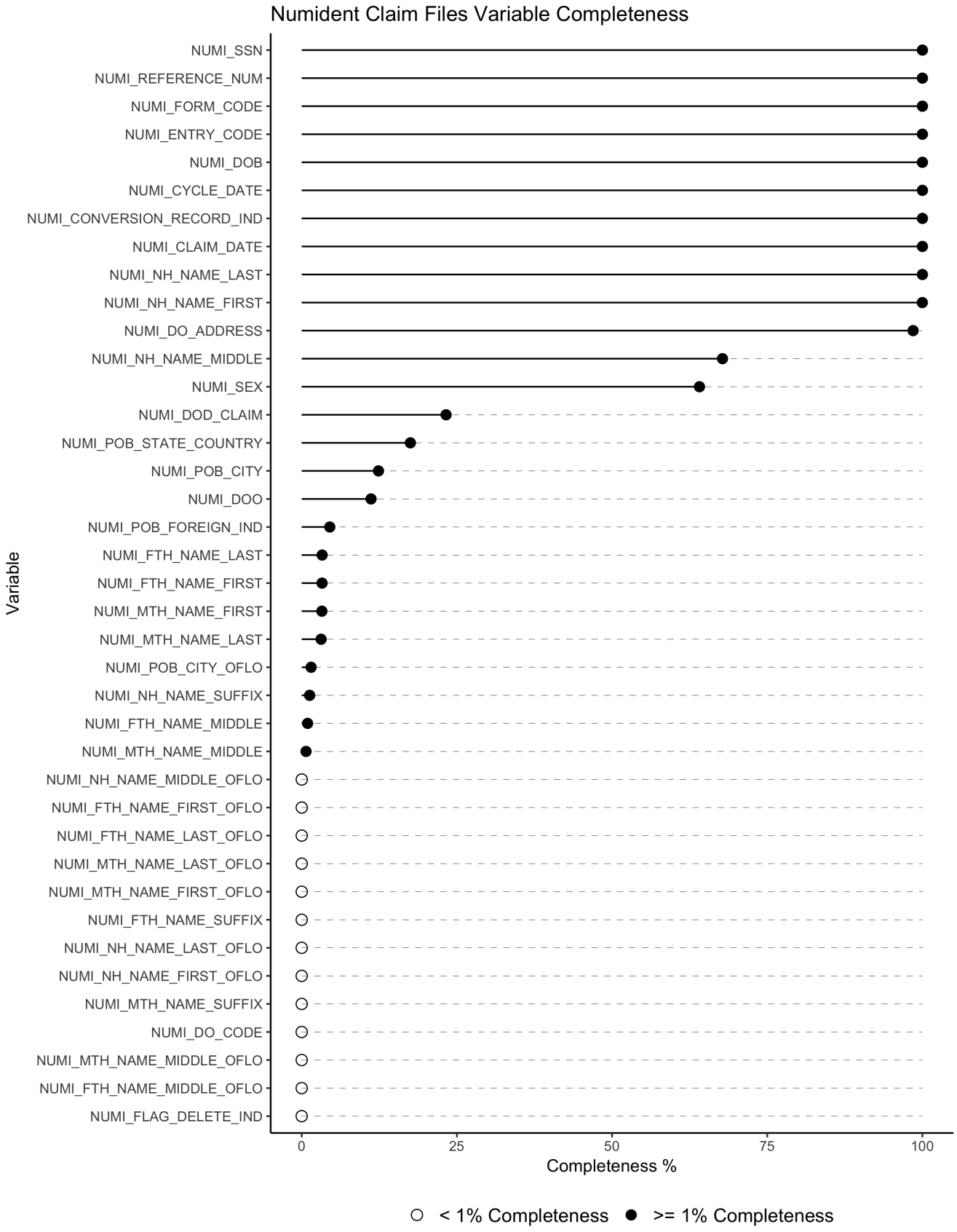

Numident Claim Files Variable Completeness
NUMI_SSN
NUMI_REFERENCE_NUM
NUMI_FORM_CODE
NUMI_ENTRY_CODE
NUMI_DOB
NUMI_CYCLE_DATE
NUMI_CONVERSION_RECORD_IND
NUMI_CLAIM_DATE
NUMI_NH_NAME_LAST
NUMI_NH_NAME_FIRST
NUMI_DO_ADDRESS
NUMI_NH_NAME_MIDDLE
NUMI_SEX
NUMI_DOD_CLAIM
NUMI_POB_STATE_COUNTRY
NUMI_POB_CITY
NUMI_DOO
NUMI_POB_FOREIGN_IND
NUMI_FTH_NAME_LAST
NUMI_FTH_NAME_FIRST
NUMI_MTH_NAME_FIRST
NUMI_MTH_NAME_LAST
NUMI_POB_CITY_OFLO
NUMI_NH_NAME_SUFFIX
NUMI_FTH_NAME_MIDDLE
NUMI_MTH_NAME_MIDDLE
NUMI_NH_NAME_MIDDLE_OFLO
NUMI_FTH_NAME_FIRST_OFLO
NUMI_FTH_NAME_LAST_OFLO
NUMI_MTH_NAME_LAST_OFLO
NUMI_MTH_NAME_FIRST_OFLO
NUMI_FTH_NAME_SUFFIX
NUMI_NH_NAME_LAST_OFLO
NUMI_NH_NAME_FIRST_OFLO
NUMI_MTH_NAME_SUFFIX
NUMI_DO_CODE
NUMI_MTH_NAME_MIDDLE_OFLO
NUMI_FTH_NAME_MIDDLE_OFLO
NUMI_FLAG_DELETE_IND
Variable
0
25
50
75
100
Completeness %
< 1% Completeness
>= 1% Completeness

Figure 15