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

Casey F. Breen † Joshua R. Goldstein‡

Draft Version: February 11, 2022

Abstract

Background: While much progress has been made in understanding the demographic determinants of mortality in the United States using individual survey data and aggregate tabulations, the lack of population-level register data is a barrier to further advances in mortality research. 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. We introduce the Berkeley Unified Numident Mortality Database (BUNMD), a cleaned and harmonized version of these records. This publicly available dataset provides researchers access to over 49 million individual-level mortality records with demographic covariates and fine geographic detail, allowing for high-resolution mortality research.

Objective: The purpose of this paper is to describe the BUNMD, discuss statistical methods for estimating mortality differentials based on this deaths-only dataset, and provide case studies illustrating the high-resolution mortality research possible with the BUNMD.

Methods: We provide detailed information on our procedure for constructing the BUNMD dataset from the most informative parts of the publicly available Social Security Numident application, claim, and death records.

Contribution: The BUNMD is now publicly available, and we anticipate these data will facilitate new avenues of research into the determinants of mortality disparities in the United States.

*For helpful discussions and feedback we thank Lynn Goodsell, Dennis M. Feehan, Leora Lawton, Ugur Yildrim, the Berkeley Human Mortality Database, and members of the CenSoc Working Group. Replication code is available on https://osf.io/eu63f/. Research reported in this publication was supported by the National Institute of Aging R01AG05894. C.F.B. was supported by National Institute of Aging T32-AG000246.

†Department of Demography, University of California, Berkeley. caseybreen@berkeley.edu.

‡Department of Demography, University of California, Berkeley. josh.goldstein@berkeley.edu.

Contents

1 Introduction

Life expectancy in the United States increased by a remarkable 30 years over the course of the 20th century. This impressive progress was driven primarily by advances in the treatment of infectious diseases and delayed mortality for those living with chronic illness (Crimmins and Zhang, 2019), but the benefits accrued unevenly. Inequality in mortality between the most advantaged and the least advantaged actually increased over time (Preston and Elo, 1995), and by the beginning of the 21st century, the gap in life expectancy between the top and bottom 1% of income earners was over 14.6 years (Chetty et al., 2016).

Despite the longstanding interest in racial and class-based inequalities in health and mortality in the United States (Schwandt et al., 2021; Elo, 2009), research is often hampered by data limitations (Card et al., 2010; Song and Coleman, 2020). Most research into the general dimensions of mortality disparities using microdata have relied on survey data, with sample sizes that preclude the analysis of smaller population subgroups such as the oldest-old. In the absence of comprehensive population-level registry data such as those found in the Scandinavian countries, researchers are increasingly turning to administrative datasets from agencies such as the Social Security Administration to answer some of the most pressing questions in social science research (Chetty et al., 2016; Card, Dobkin and Maestas, 2008; Card et al., 2010; Meyer and Mittag, 2019; Ruggles, 2014). While the recent introduction of the United States Mortality Database has created a valuable new resource for studying aggregate mortality trends at the state level (USMDB, 2021), public-use administrative microdata for mortality research remains scarce. When available, it is often cumbersome to use.

We introduce the Berkeley Unified Numident Mortality Database (BUNMD), a cleaned and harmonized version of administrative mortality records from the Social Security Administration.[1] The BUNMD represents one of the first publicly available, large-scale administrative microdata resources for studying mortality. We anticipate that the size (N = 49 million) and spatial detail will open up new avenues for high-resolution mortality research. Furthermore, the open-access nature of the dataset will ensure this research is reproducible and extendable.

1.1 Background

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 date of birth, date of death (if applicable), claims status, and other background information such as birthplace, race, sex, parents' first and last names, and ZIP Code of residence at the time of death. An internal, restricted-access version of the Numident has been used to study mortality by researchers either employed by the Social Security Administration (Waldron, 2007) or collaborating with Social Security researchers (Mehta et al., 2016; Elo et al., 2004).[2] Until recently, opportunities to study these data were not available to the general research community.

In 2013, the Social Security Administration transferred a large portion of their Numident records to the National Archives and Records Administration (NARA). The NARA public release of these records in 2019, which we term "NARA Numident," created one of the first

[1] The BUNMD can be downloaded here: https://censoc-download.demog.berkeley.edu/

[2] Researchers can apply to access the restricted-use version of the Numident ("Census Numident") in one of 32 Federal Statistical Research Data Centers (Finlay and Genadek, 2021). This version of the Numident can also be linked internally to other Census data, such as the 2000 or 2010 decennial census.

Variable	Description	Numident Source
ssn	Social Security Number	Death Entry
fname	First name	Death Entry
mname	Middle name	Death Entry
lname	Last Name	Death Entry
byear	Year of birth	Death Entry
bmonth	Month of birth	Death Entry
bday	Day of birth	Death Entry
dyear	Year of death	Death Entry
dmonth	Month of death	Death Entry
dday	Day of death	Death Entry
zip_residence	ZIP Code of residence at death	Death Entry
sex	Sex	Death, Application, or Claim Entry
race_first	Race (first)	Application Entry
race_last	Race (last)	Application Entry
bpl	Place of birth	Application or Claim Entry
father_fname	Father's first name	Application or Claim Entry
father_mname	Father's middle name	Application or Claim Entry
father_lname	Father's last name	Application or Claim Entry
mother_fname	Mother's first name	Application or Claim Entry
mother_mname	Mother's middle name	Application or Claim Entry
mother_lname	Mother's last name	Application or Claim Entry
race_change	Change of race	Constructed
death_age	Age of death (years)	Constructed
socstate	State in which SS card issued	Constructed
age_first_app	Age of first application	Constructed
number_apps	Total number of applications	Constructed
number_claims	Total number of claims	Constructed
weight	Weight variable	Constructed
ccweight	Complete case person-weight	Constructed

Table 1: Variables in the BUNMD.

public-use administrative mortality datasets in the U.S., offering nearly complete coverage for those 65+ dying between 1988 and 2005. The original collection of NARA Numident records were formatted as a set of 60 separate fixed-width text files with over 150 different fields. We introduce a cleaned and harmonized version of the NARA Numident records: the Berkeley Unified Numident Mortality Database (BUNMD). The BUNND file condenses the Numident death, application, and claim records into a single user-friendly file with one record per person, including over 49 million death records and the 30 variables displayed in

Table 1. We anticipate the public availability of the BUNMD will open up new avenues of research using administrative records.

The BUNMD offers several advantages for the study of mortality. First, the large sample size enables the comparison of birth cohorts, small population subgroups, and small geographic areas. Second, the nearly complete death coverage for those 65+ allows for the study of mortality disparities at the oldest ages, when cohorts have only a few remaining survivors. Finally, the public nature of the BUNMD means that it can be linked – either through Social Security number or a combination of identifiers such as first name, last name, year of birth, and place of birth – to other datasets with covariates of mortality.

We have linked the BUNMD records to the complete count 1940 Census to create a publicly-available linked administrative dataset for the study of mortality (Goldstein et al., 2021)[3]. We established matches using the ABE exact record linkage algorithm, which requires an exact match on first name, last name, and place of birth, but allows for flexibility ±2 years on birth year (Abramitzky et al., 2019). The resulting dataset, termed CenSoc-Numident, contains over 7.9 million records and allows researchers to take advantage of the rich measures available in the public 1940 Census: education, geography, home-ownership, income, occupation, place of birth, family structure, and parental birthplace. Researchers can also work with a restricted version of the 1940 Census, allowing access to additional measures such as full names and exact street-level addresses.[4]

[3]To download CenSoc datasets and access tutorials for working with the data, please visit https://censoc.berkeley.edu/

[4]We have also linked the 1940 Census to a public version of the Death Master File (DMF). While the public 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). The resulting dataset, the CenSoc-DMF, has a larger window of high death coverage (1975-2005) but is restricted only to men. Both the CenSoc-DMF and CenSoc-Numident have been further linked to World War II enlistment records, allowing for the investigation of army rank, height, weight, and other administrative variables.

The BUNMD records pose a unique challenge for mortality estimation because the BUNMD only has high death coverage for a left and right ("doubly") truncated window of 1988-2005. In addition, because the dataset does not include survivors, researchers must use statistical methods that rely on the distribution of deaths by age within cohorts. We illustrate those methods below.

The remainder of the paper proceeds as follows. In Section 2, we describe the structure and content of the public NARA Numident records and our procedure for combining the records into a cleaned and harmonized file with a single record per person. In Section 3, we discuss statistical methods for estimating mortality differentials based on this deaths-only dataset and provide examples of their use. We conclude in Section 4 with several substantive case studies demonstrating the high-resolution mortality research possible with the BUNMD.

2 The Structure and Content of the NARA Numident Records

The NARA Numident consists of three different types of records: applications (SS-5), claims, and deaths. Each individual in the Numident may have an application, claim, and/or death record; each application and claim record can contain multiple entries. Records can contain multiple entries because the Social Security Administration adds a new entry to the Numident when a Social Security cardholder submits a new application or claim (Record Group 47, 2019).

As shown in Table 2, the NARA Numident contains 49.5 million death record entries

containing full name, social security number, sex, date of birth, and date of death. Additionally, it contains 72.1 million application (SS-5) entries for 40.8 million unique individuals. The application entries contain information extracted from applications for a Social Security card and applications for Social Security account number, including: full name, race, sex, birthplace, date of birth, parents' full names, social security number, and other administrative information. Finally, the NARA Numident contains 25.2 million claim entries for 25.1 million unique individuals. The information in the claim entries is largely redundant with that of the application and death records: full name, social security number, date of birth, sex, and type of claim.

To further illustrate the structure and content of the NARA Numident records, Table 3 shows 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, the NARA Numident contains one application and one death record. For Lana Turner, the NARA Numident contains one death record and four different application records, corresponding to name changes each time she got married.

Record Type	Total Entries	Total Records (Persons)	Entries per Person
Death	49,459,293	49,459,293	1.000
Applications	72,120,516	40,870,455	1.765
Claims	25,228,257	25,140,847	1.004

Table 2: Number of records and entries in the NARA Numident. A Numident application or claim record may contain more than one entry.

We followed three steps to create the cleaned and harmonized BUNMD from the original NARA Numident records: (1) we selected names and birth and death dates from the death entries; (2) we added key covariates from the application and claim entries, using a set of

Table 3: 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_residence**	**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 date**				**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_residence**	**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.

decision rules to reconcile discrepant values across entries[5]; (3) we constructed new variables reporting total number of applications, total number of claims, age at first Social Security application, and state in which the Social Security number was issued. See Appendix Section B for technical details.

2.1 BUNMD Mortality Coverage

To assess the mortality coverage of the BUNMD, we compare death counts to the gold-standard Human Mortality Database (HMD) (HMD, 2021). While the universe for the HMD differs slightly from the BUNMD, both datasets capture nearly all deaths occurring in the U.S., making the HMD an excellent benchmark for the BUNMD deaths. The HMD excludes deaths to non-residents and deaths to U.S. residents living outside of the U.S., either in foreign countries or in territories (HMD, 2021). The BUNMD excludes deaths to individuals without a Social Security record but captures deaths to Social Security Number holders dying abroad (Record Group 47, 2019).

Figure 1 compares the total number of deaths for persons age 65+ (when coverage is highest) in the BUNMD to the HMD. Death coverage is nearly complete between 1988 and 2005 but drops dramatically to less than 10% coverage outside of this window.[6] Figure 2 shows the coverage visualized on an age-period Lexis surface (Schöley and Willekens, 2017). Each "cell" represents death coverage for a given age and year. Death coverage is defined as

[5]In order to study name changes, shifts in racial self-identity, and other features, the original NARA Numident records are useful and are available upon request.

[6]The exact data generating process – why some Numident records were transferred from the Social Security Administration to NARA for public release and not others – is unclear. 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.

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. Death coverage is highest for those age 65+ dying between 1988 and 2005. For those dying before age 65, coverage is generally between 50% and 75%.

2.2 BUNMD Samples and Weights

We created two weighted BUNMD samples. Sample 1 includes birth cohorts from 1900 to 1940 who died after age 65+ between 1988 to 2005. Sample 2 – the "complete case" sample – is the subset of Sample 1 records with complete information on sex, birthplace, and race. For both samples, we constructed a set of person-level weights to make the distributions of deaths in the BUNMD match aggregate population totals. Specifically, we broke the sample into cells cross-classified by year of birth, year of death, age at death, and sex. We assigned every person in a given cell a weight equal to the ratio of count of deaths in the HMD to count of deaths in the BUNMD sample:

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

3 Mortality Estimation without Denominators

While the BUNMD includes deaths from 1900-2007, we recommend researchers restrict their analyses to deaths occurring in the high death coverage of 1988-2005. Restricting to this window is important for two reasons. First, when death coverage is low, the death records available are highly select – that is, these deaths were not included in the BUNMD at

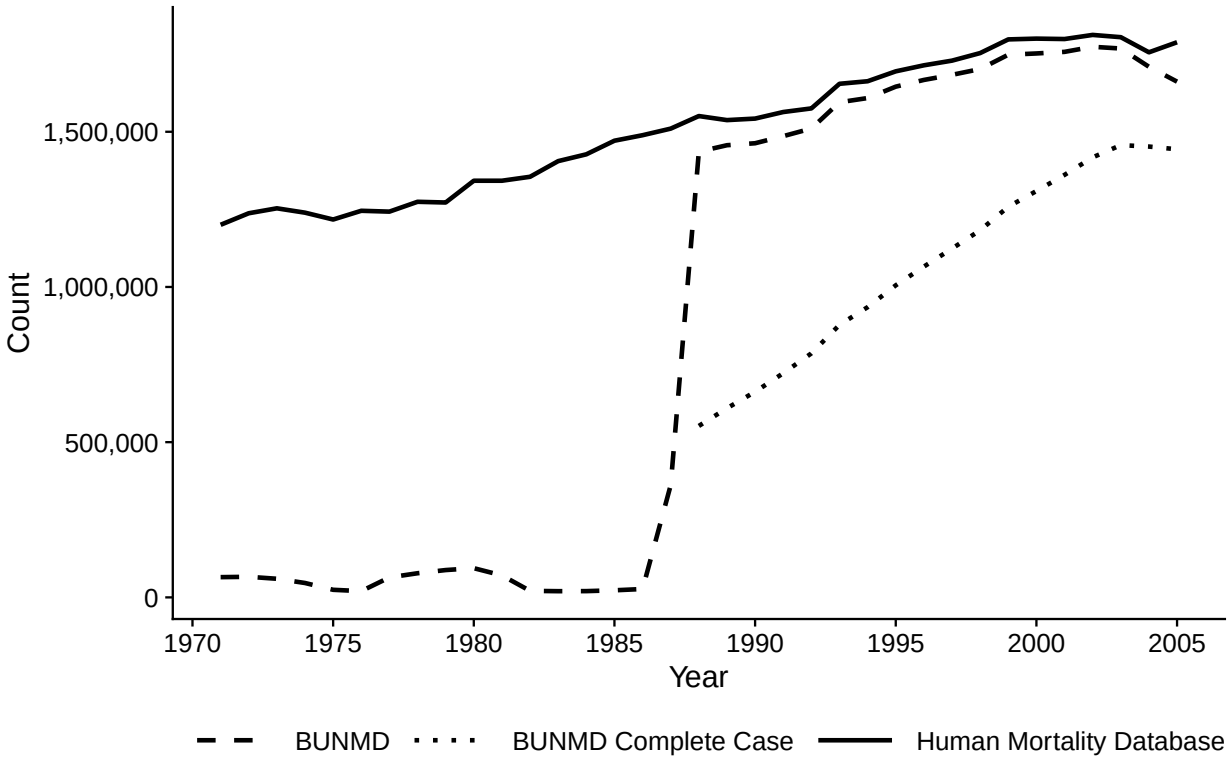

Figure 1: BUNMD Death Coverage for persons 65+. Coverage is most complete in the BUNMD for the window of 1988 to 2005. The complete cases include information on sex, birthplace, and race.

random. This is not an issue for the high coverage death window, where death coverage is over 95%+. Second, not restricting to this high coverage death window will preclude the use of any parametric methods that rely on distribution of deaths within a cohort.

When death coverage is restricted to the window of 1988 and 2005, the mortality records are doubly truncated and have no information on denominators (Alexander, 2018*a*). Conventional survival analysis tools cannot be applied in this setting: calculating mortality rates, survival functions, and hazard functions all require some measure of exposure to risk – a denominator. Furthermore, methods to describe the effect of covariates on mortality, such as cox-proportional hazards models and OLS linear regression on age at death, are biased in the presence of double truncation (Lin and Wei, 1989; Rennert and Xie, 2018).

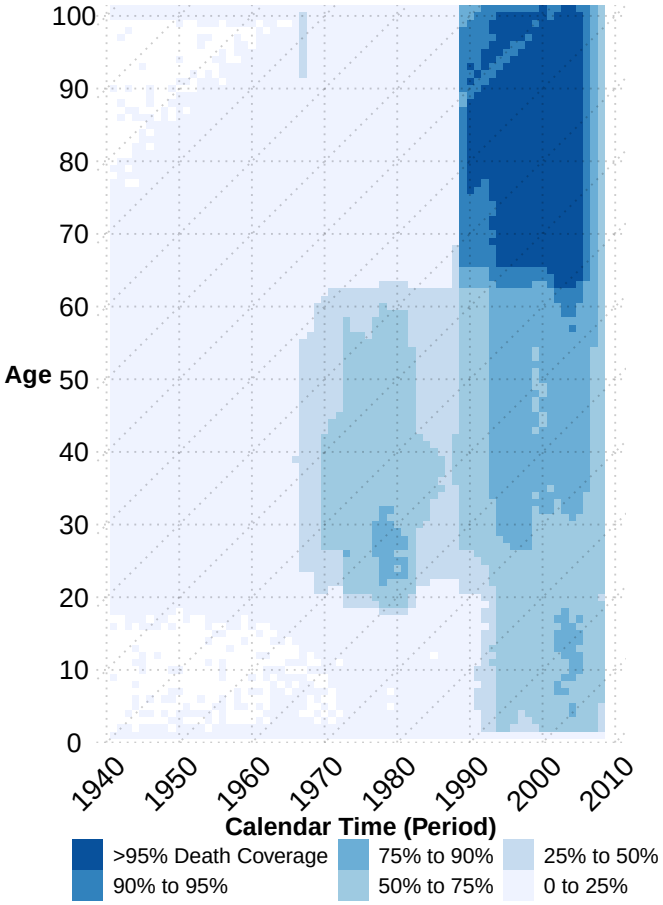

Figure 2: Lexis diagrams of BUNMD death coverage for 1940 to 2010. Coverage is highest for those age 65+ dying between 1988 and 2005.

To illustrate the effect of double truncation on estimated mortality differentials, Figure 3 shows simulated death distributions for two subpopulations. Panel (a) shows the untruncated distributions of deaths after age 65 for two subpopulations. The blue subpopulation has a mean age of death of 83 years and the red subpopulation has a mean age of death of 80 years: a difference of 3 years. Panel (b) shows the exact same distributions of death doubly truncated to include deaths from age 78 to 95 (the observation window for the BUNMD cohort of 1910). Under double truncation, the blue subpopulation has a mean age of death of 84.9 years and the red subpopulation has a mean age of death of 86 years: a difference of only 1.1 years. In this example, the double truncation will understate the true difference

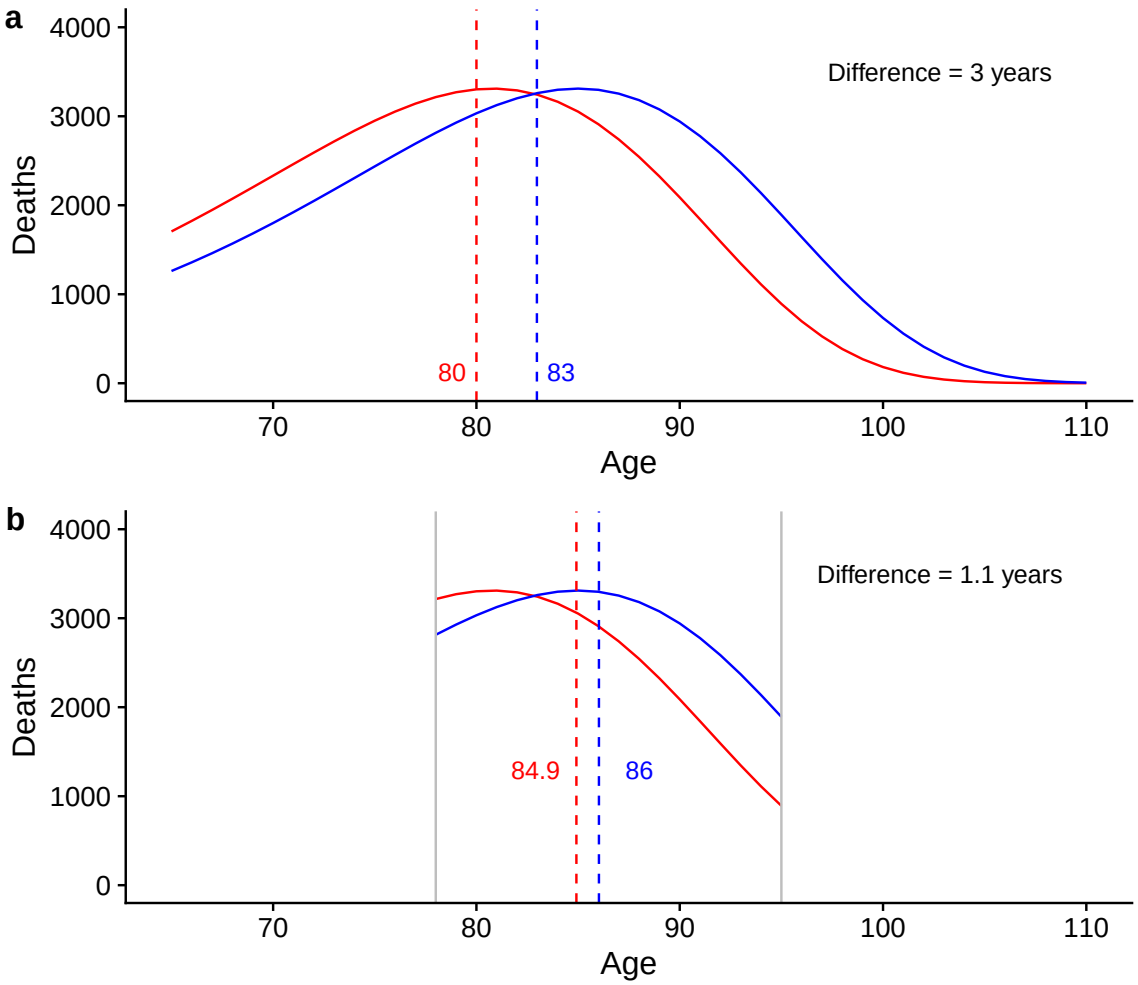

Figure 3: Simulated distributions of age of death for two populations. Gray lines represent hypothetical left and right truncation. The difference in mean age of death is 3 years for the untruncated distributions (panel a) and 1.1 years for the truncated distribution (panel b) – the truncated difference understates the true difference by a factor of 2.72.

in mean age of death by a factor of 2.72. Broadly, the narrower the age window considered, the more understated this difference will be.

To address this challenge, we have developed two different methods,[7] which can be chosen based on suitability for the research question of interest. The first method is to use OLS

[7]For extinct cohorts (in which all members have died), it is possible to use classical methods of "extinct generations" to calculate mortality rates. Specifically, the total number of survivors at a given age can be found by summing up all the deaths occurring above that age, and then the age-specific mortality rates can be estimated from the age-specific ratios of deaths to survivors. However, these methods are only appropriate for the cohorts born before 1900, for which only a few survivors to age 105 will die after 2005.

regression on age at death with birth year fixed effects. The second method is to fit parametric Gompertz survival models, using maximum likelihood estimation (MLE) to explicitly account for the double truncation.

The Gompertz model is the oldest and most prominent model to describe the characteristic pattern of older age human mortality. It assumes mortality hazards rise exponentially with age

$$h(x) = ae^{bx} \tag{2}$$

where $h(x)$ is the hazard at age x, a is the baseline level of mortality, and b is the rate at which mortality increases with age. The Gompertz model can also be reparameterized in terms of modal age at death

$$h(x) = be^{b(x-M)} \tag{3}$$

where M is the modal age at death, the age at which the highest number of deaths occur. This reparameterization has two key advantages: the modal age M parameter has a more intuitive explanation than a, and M and b are less highly correlated than a and b, which is beneficial for model fitting (Missov et al., 2015; Alexander, 2018*b*).

3.1 Method 1: Linear Regression on Age of Death

Ordinary Least Squares (OLS) linear regression on age of death is an easy and effective way to analyze the BUNMD mortality records. The regression coefficients report the association between covariates and the mean age at death. The key limitation to this approach is that

it will produce biased estimates in the presence of double truncation (Alexander, 2018*a*; Greene, 2005; Dörre and Emura, 2019). Models of the form:

$$\underbrace{D}_{\text{Age at Death}} = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\gamma_t t}_{\substack{\text{Birth Year} \\ \text{Fixed Effects}}} + \underbrace{\mathbf{X}\beta}_{\substack{\text{Covariates and} \\ \text{Design Matrix}}} + \underbrace{\epsilon}_{\text{Error Term}} \tag{4}$$

provide estimates of the association of the covariates on the age of death in the sample, controlling for birth cohort truncation effects. The regression model must be fit with fixed effects for year of birth because the left and right truncation ages vary by birth cohort. For instance, in the BUNMD we observe the cohort of 1900 dying between ages 87 and 105 and observe the cohort of 1920 dying between ages 67 and 85.

In sum, researchers can do their full statistical analysis, including hypothesis testing and model selection, using OLS regression on age at death with dummy variables for year of birth. However, the magnitude of any coefficients will generally be attenuated (biased towards 0) by the double truncation in this setting.[8]

3.2 Method 2: Parametric Gompertz Survival Models (for Truncated Data)

Our second method uses maximum likelihood methods to estimate age-specific mortality with multivariate predictors, assuming the distribution of the age at death among members

[8]For insight into this attenuation, suppose that high school graduates live on average 10 years longer than those with less education. If we only observe deaths between the ages of 70 and 75, the gap in life expectancy is necessarily compressed (it must be less than 6 years). All else equal, the estimated OLS regression coefficient on a dichotomous predictor is the average difference in the dependent variable between reference category (no high school) and comparison group (completed high school). Thus, a regression estimating the association between completing high school and age of death will estimate an attenuated (biased towards 0) regression coefficient for the high school predictor.

of a cohort follows a parametric Gompertz model. This method adapts the usual method of parametric survival analysis to our specific case of observing only individuals who have died during a doubly truncated window. For any parametric model, we define a likelihood given the deaths we are able to observe. For doubly truncated cohorts, with known left truncation a and known right truncation b, we can define the conditional distribution of the age at death among members of a cohort to be

$$f_{trunc} = \frac{f_\theta(x)}{\int_a^b f_\theta(x)dx} = \frac{f_\theta(x)}{F_\theta(b) - F_\theta(a)} \tag{5}$$

with likelihood of

$$L(\theta|X) = \prod \frac{f_\theta(x_i)}{F_\theta(b) - F_\theta(a)} \tag{6}$$

This parametric Gompertz distribution can be extended to include covariates by making the "proportional hazards" assumption that each covariate has a direct multiplicative effect on the hazard rate. The estimated hazard for individual i at age x with parameters β is given by:

$$h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \tag{7}$$

t

where $h(x)$ is the hazard at age x, a_0 is some baseline level of mortality, b_0 gives the rate of increase of mortality, Z_i are the covariates for person i (e.g., years of education, place of birth), and β is the set of parameters. The estimates of the vector β of parameters

can be obtained by maximizing the likelihood, or, equivalently, the log-likelihood. Code for implementing this parametric Gompertz approach is available in the **gompertztrunc** package for the R statistical programming language.

This approach has several advantages. First, researchers can use either standard optimization routines to maximize the likelihood as provided in the **gompertztrunc** package or extend this approach to analogous Bayesian methods. Second, it can be adopted to include other parametric models of old-age mortality, such as the Makeham model, which includes an additional parameter to allow for mortality deceleration in the oldest ages. Finally, researchers can extend this model to allow for variation in the underlying Gompertz parameters across cohort and population subgroups using a hierarchical structure. For instance, a researcher may want to allow the Gompertz slope to vary over time.

It is also 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*a*). This approach 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.

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 immigrants typically face persistent socioeconomic disadvantage in the U.S. This phenomenon, termed the

"immigrant paradox," has long been observed for Hispanic immigrants (Hummer et al., 2000; Fenelon, Chinn and Anderson, 2017). Accurate estimates of mortality differentials for many other immigrant groups have typically not been possible due to insufficient sample sizes. Recently, Mehta et al. (2016) used 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 the Mehta et al. (2016) findings using publicly available data. In our analysis, we restrict to foreign-born individuals who applied for Social Security cards before age 65 and before the year 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."

4.1.1 Country-of-Birth Differences

To investigate heterogeneity in mortality experience by country of origin, we examine the mortality differences among immigrants from the top 18 sending countries, restricting to the birth cohorts of 1910 to 1919. We use the Gompertz parametric MLE approach introduced in Section 3 to estimate the difference in life expectancy at age 65 (e_{65}), fitting models using weights separately for men and women. Figure 4 plots the estimated difference in mean age of death between native-born and the foreign-born.

The BUNMD's large sample size gives us enough precision to investigate heterogeneity in mortality advantage by individual sending country. While Mehta et al. (2016) reported results for broader regions (e.g., Central America, western/eastern Europe, and Africa),

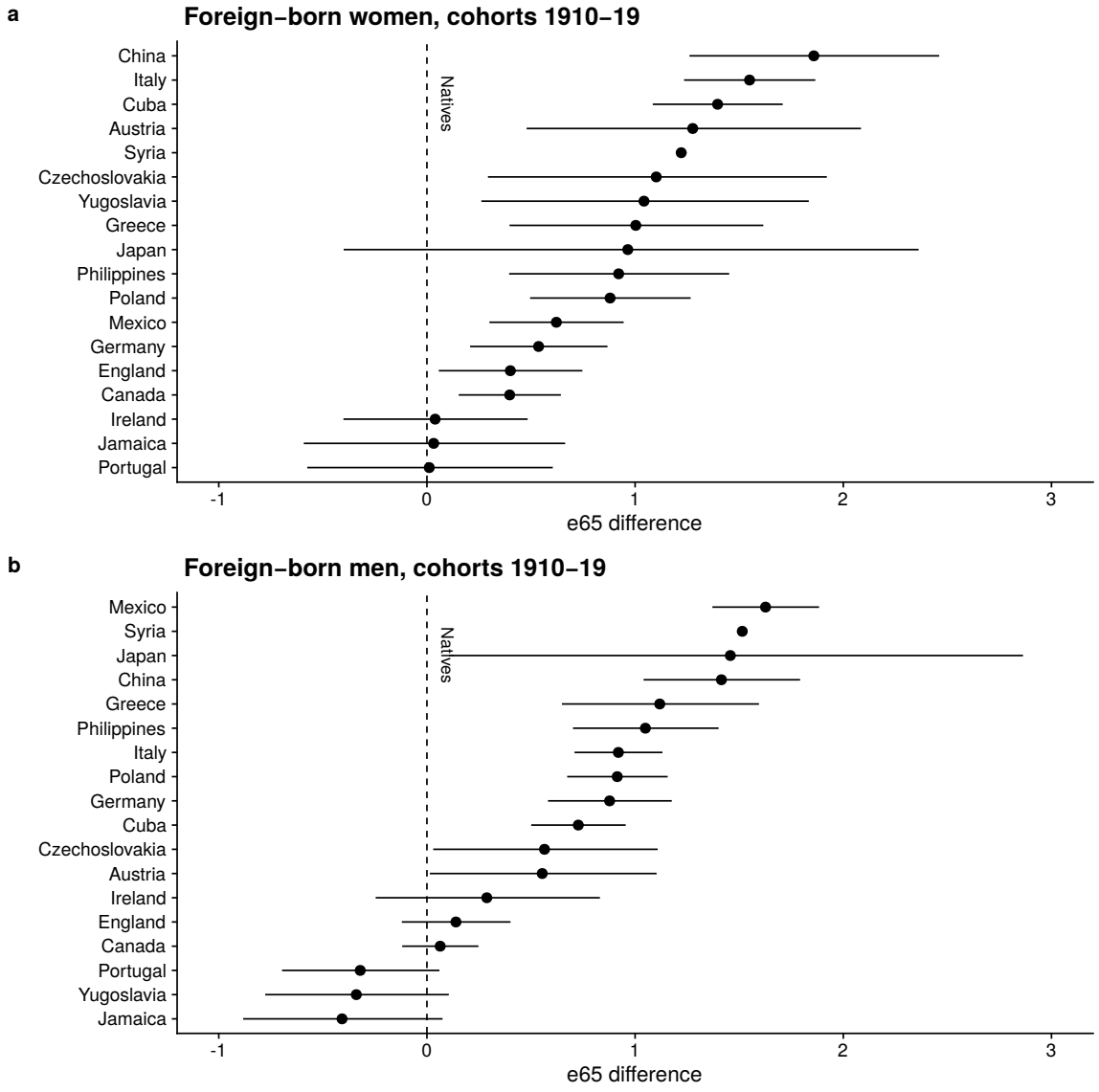

Figure 4: Difference in e_{65} between natives and foreign-born populations.

finding immigrants from every region had lower mortality than natives, we are able to conduct a country-level analysis. Our estimates show not all immigrants to the U.S. have a mortality advantage – Jamaican, Portuguese, and Yugoslavian men suffer 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 notable 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 BUNMD. 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)

4.2 Geographic Variation in Mortality by Place of Death

Geographic disparities in mortality have been well-documented in the U.S. The importance of state-level context has been investigated through the lenses of socioeconomic status (Montez et al., 2019), income inequality (Subramanian and Kawachi, 2004), and geographic exposures early in life (Xu et al., 2020). The gap in life expectancy persists into older ages: e_{50} is 3.4 years higher in Minnesota, the state with the highest longevity, than Mississippi, the state

with the lowest longevity (Wilmoth, Boe and Barbieri, 2010). Less is known about geographic disparities in mortality in the U.S. at finer geographic levels.

The BUNMD includes the 9-digit ZIP Code of residence at the time of death for approximately 70% of records, allowing the direct estimation of small-area mortality rates. To illustrate the BUNMD's potential for small-area mortality estimation, we investigate geographic differences in Ohio's Cuyahoga County, which surrounds the city of Cleveland. Cuyahoga County is noteworthy because of its history of within-county income inequality (Tumin et al., 2018) and racial segregation (Tomer, 2020).

Figure 5, panel (a) plots differences in e_{65} by ZIP Code for the birth cohorts of 1910-1919 in Ohio's Cuyahoga County, estimated using the parametric Gompertz MLE method. The black boundary indicates the city of Cleveland. Life expectancy is lower in inner-city Cleveland, and higher in its surrounding affluent suburbs, reflecting the high level of racial and class-based segregation.

This estimation procedure using the parametric Gompertz MLE method is based on the assumption of a closed population. In populations where there is migration (e.g., states, counties, zip codes), the method can still work but it assumes that there is not age-selective net migration (Bureau, 2003).

4.3 Linking to other datasets

The BUNMD can be linked to other individual-level datasets using dates of birth, full names, or Social Security number. Contextual data can also be merged with the BUNMD on ZIP Code (or state) of residence at time of death. To illustrate, we have linked contextual

data on the amount of Social Security Benefits paid in each ZIP Code obtained from a report on the Master Beneficiary Record, the principal administrative file of Social Security beneficiaries (Administration, 2005).

Figure 5, panel (b) shows average monthly Social Security benefits per eligible retired worker by ZIP Code in 2005. As the amount of benefits a qualified retiree receives increases with total lifetime earnings, average Social Security retirement benefits per retiree can be used as a measure of a ZIP Code's wealth. A comparison of panel (a) and panel (b) shows that higher average monthly Social Security benefits per retiree is strongly associated with a higher e_{65} at the ZIP Code level.

5 Considerations for researchers

There are several caveats and limitations of the BUNMD that warrant discussion. First, as with all administrative data, the original Numident records were intended for a specific administrative purpose and have been repurposed for mortality research. One implication of this is that the exact data generating process – why some Numident records were chosen for public release by the Social Security administration and not others – is unknown. We recommend researchers restrict their analysis to deaths occurring in the window of 1988-2005 when death coverage is 95%+ complete; this minimizes the risk of introducing selection bias by including highly-select deaths occurring in the low-mortality coverage window.

While the BUNMD includes weights to account for incomplete mortality coverage, these weights do not account for different probabilities of inclusion in the BUNMD for population subgroups. We recommend that researchers check the probability of being included in the

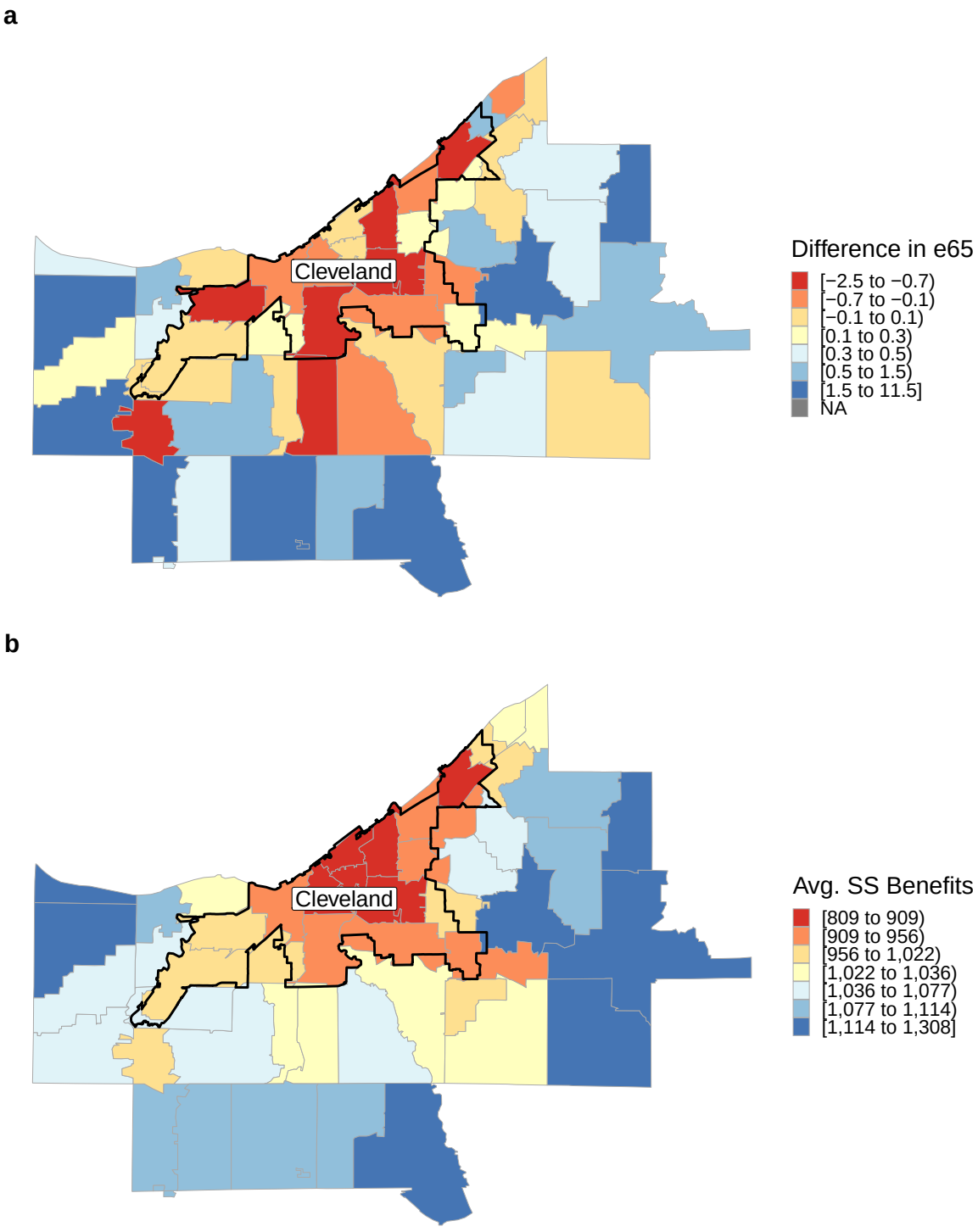

Figure 5: **Mortality and wealth in Cuyahoga County**. Panel (a) shows the difference in e_{65} in Cuyahoga County for the birth cohorts of 1910-1919 by ZIP Code of residence at time of death. Panel (b) shows the average monthly Social Security benefits in U.S. dollars by ZIP Code. The black line shows the city boundary for Cleveland.

BUNMD is independent of cohort within a given time period. Additionally, each variable in the BUNMD has a different proportion of missing values. For instance, information on ZIP Code of residence at time of death is only available for 70% of cases (see Appendix Section B.3 for a complete list of each variable's proportion of missing values). Researchers doing analyses conditional on the availability of such variables should confirm this does not bias the respresentativity of the sample they are using for analysis.

Researchers studying certain population subgroups, such as immigrants, must be careful to account for people entering into the BUNMD during the observation window. If people arrive in the middle of the observation period, the distribution of deaths observed will be upwardly biased. For instance, a person gaining access to the Social Security Administration in 2003 would only be observed for the years of In these settings, we recommend researchers restrict to cases where the first application record was submitted before 1988. Because the Numident death records capture Social Security card holders dying outside the U.S., there is less risk of people leaving during the observation window.

For some individuals in the BUNMD, certain measures may have been reported several times in the original NARA Numident records. For example, sex may be reported on several application entries for a given individual. While we used a principled set of selection rules to choose the value ultimately included in the BUNMD, reporting or transcription errors will inevitably remain. For researchers studying name changes or shifts in racial self-identification, the original NARA Numident records are available upon request. Additionally, the original records contain an uncleaned 12-character string variable reporting "city and/or county of birth" that is not released in the BUNMD.

Finally, the ethno-racial categories on the Social Security application changed in 1980.

Before 1980, the application form had three options: (1) White, (2) Black, and (3) Other. After 1980, the form gave five options: (1) White, (2) Black, (3) Asian, Asian American, or Pacific Islander, (4) Hispanic, and (5) North American Indian or Alaskan Native. For individuals with multiple application entries, the responses to the race question was most stable for Blacks (99%) and Whites (98%), and least stable for Asians (92%) and American Indian or Alaskan Native (85.1%) (Breen, 2021). The BUNMD includes the ethno-racial identity reported on both the first and last submitted application entry, the dates of first and last application entry, and a flag variable indicating whether a shift in racial self-identification had occurred.

6 Conclusion

The BUNMD, with over 49 million death records, is a novel data resource for mortality researchers, especially those interested in old age mortality and smaller population subgroups. Using the statistical methods described in this paper, many more researchers will be able to conduct high-resolution mortality research, furthering our understanding of the dimensions of mortality disparities in the U.S. Furthermore, because the BUNMD is an open-access resource, it vastly increases opportunities for mortality research that is reproducible and extendable.

References

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James Feigenbaum and Santiago Pérez. 2019. Automated Linking of Historical Data. Technical Report w25825 National Bureau of Economic Research Cambridge, MA: .

Administration, Social Security. 2005. "OASDI Beneficiaries by State and ZIP Code, 2005." p. 726.

Alexander, Monica. 2018*a*. Deaths without Denominators: Using a Matched Dataset to Study Mortality Patterns in the United States. Preprint SocArXiv.

Alexander, Monica. 2018*b*. "Gompertz Mortality Models." https://www.monicaalexander.com/posts/2018-02-15-gompertz/.

Barron, Erma. 1982. "Meaning of the Social Security Number." p. 2.

Breen, Casey. 2021. Shifts in Racial Self-Identification for the Greatest Generation: Evidence from Social Security Administrative Data. Preprint SocArXiv.

Bureau, Census. 2003. "Internal Migration of the Older Population: 1995 to 2000.".

Card, David, Carlos Dobkin and Nicole Maestas. 2008. "The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare." *American Economic Review* 98(5):2242–2258.

Card, David E., Raj Chetty, Martin S. Feldstein and Emmanuel Saez. 2010. "Expanding Access to Administrative Data for Research in the United States." *SSRN Electronic Journal* .

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.

Crimmins, Eileen M. and Yuan S. Zhang. 2019. "Aging Populations, Mortality, and Life Expectancy." *Annual Review of Sociology* 45(1):69–89.

Dörre, Achim and Takeshi Emura. 2019. *Analysis of Doubly Truncated Data: An Introduction*. SpringerBriefs in Statistics Singapore: Springer Singapore.

Elo, Irma T. 2009. "Social Class Differentials in Health and Mortality: Patterns and Explanations in Comparative Perspective." *Annual Review of Sociology* 35:553–572.

Elo, Irma T., Cassio M. Turra, Bert Kestenbaum and B. Reneé Ferguson. 2004. "Mortality among Elderly Hispanics in the United States: Past Evidence and New Results." *Demography* 41(1):109–128.

Fenelon, Andrew, Juanita J. Chinn and Robert N. Anderson. 2017. "A Comprehensive Analysis of the Mortality Experience of Hispanic Subgroups in the United States: Variation by Age, Country of Origin, and Nativity." *SSM - Population Health* 3:245–254.

Finlay, Keith and Katie R. Genadek. 2021. "Measuring All-Cause Mortality With the Census Numident File." *American Journal of Public Health* 111(S2):S141–S148.

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.

Goldstein, Joshua .R, Monica Alexander, Casey Breen, Andrea Miranda-González, Felipe Menares, Osborn, Maria and Yildrim, Ugur. 2021. "CenSoc Mortality File: Version 2.0.".

Greene, William H. 2005. "Censored Data and Truncated Distributions." *SSRN Electronic Journal* .

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.

HMD. 2021. "Human Mortality Database." *Available at www.mortality.org or www.humanmortality.de* .

Hummer, Robert A., Richard G. Rogers, Sarit H. Amir, Douglas Forbes and W. Parker Frisbie. 2000. "Adult Mortality Differentials among Hispanic Subgroups and Non-Hispanic Whites." *Social Science Quarterly* 81(1):459–476.

Lin, D. Y. and L. J. Wei. 1989. "The Robust Inference for the Cox Proportional Hazards Model." *Journal of the American Statistical Association* 84(408):1074–1078.

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):1109–1134.

Meyer, Bruce D. and Nikolas Mittag. 2019. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net." *American Economic Journal: Applied Economics* 11(2):176–204.

Missov, Trifon I., Adam Lenart, Laszlo Nemeth, Vladimir Canudas-Romo and James W. Vaupel. 2015. "The Gompertz Force of Mortality in Terms of the Modal Age at Death." *Demographic Research* 32:1031–1048.

Montez, Jennifer Karas, Anna Zajacova, Mark D. Hayward, Steven H. Woolf, Derek Chapman and Jason Beckfield. 2019. "Educational Disparities in Adult Mortality Across U.S. States: How Do They Differ, and Have They Changed Since the Mid-1980s?" *Demography* 56(2):621–644.

Preston, Samuel H. and Irma T. Elo. 1995. "Are Educational Differentials in Adult Mortality Increasing in the United States?" *Journal of Aging and Health* 7(4):476–496.

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

Rennert, Lior and Sharon X. Xie. 2018. "Cox Regression Model with Doubly Truncated Data: Cox Regression Model with Doubly Truncated Data." *Biometrics* 74(2):725–733.

Ruggles, Steven. 2014. "Big Microdata for Population Research." *Demography* 51(1):287–297.

Schöley, Jonas and Frans Willekens. 2017. "Visualizing Compositional Data on the Lexis Surface." *Demographic Research* 36:627–658.

Schwandt, Hannes, Janet Currie, Marlies Bär, James Banks, Paola Bertoli, Aline Bütikofer, Sarah Cattan, Beatrice Zong-Ying Chao, Claudia Costa, Libertad González, Veronica Grembi, Kristiina Huttunen, René Karadakic, Lucy Kraftman, Sonya Krutikova, Stefano Lombardi, Peter Redler, Carlos Riumallo-Herl, Ana Rodríguez-González, Kjell G. Salvanes, Paula Santana, Josselin Thuilliez, Eddy van Doorslaer, Tom Van Ourti, Joachim K. Winter, Bram Wouterse and Amelie Wuppermann. 2021. "Inequality in Mortality between Black and White Americans by Age, Place, and Cause and in Comparison to Europe, 1990 to 2018." *Proceedings of the National Academy of Sciences* 118(40):e2104684118.

Song, Xi and Thomas S Coleman. 2020. "Using Administrative Big Data to Solve Problems in Social Science and Policy Research." p. 19.

Subramanian, S. V. and Ichiro Kawachi. 2004. "Income Inequality and Health: What Have We Learned So Far?" *Epidemiologic Reviews* 26(1):78–91.

Tomer, Lara Fishbane and Adie. 2020. "How Cleveland Is Bridging Both Digital and Racial Divides.".

Tumin, Dmitry, Michelle Menegay, Emily A. Shrider, Michael Nau and Rachel Tumin. 2018. "Local Income Inequality, Individual Socioeconomic Status, and Unmet Healthcare Needs in Ohio, USA." *Health Equity* 2(1):37–44.

USMDB. 2021. "United States Mortality DataBase." *University of California, Berkeley* .

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.

Wilmoth, John R., Carl Boe and Magali Barbieri. 2010. *Geographic Differences in Life Expectancy at Age 50 in the United States Compared with Other High-Income Countries*. National Academies Press (US).

Xu, Wei, Michal Engelman, Alberto Palloni and Jason Fletcher. 2020. "Where and When: Sharpening the Lens on Geographic Disparities in Mortality." *SSM - Population Health* 12:100680.

Supplemental Information

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

A 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 original NARA Numident records contain three types of records: application, claim, and death; each set of records was packaged separately as a set of 20 fixed-width .txt files (3 x 20 = 60 files in total).

A.1 NARA Numident Metadata

We obtained three documents from the National Archives Technical Documentation series[9]:

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

The record layout documents contain variable descriptions, value labels, technical notes,

[9]We obtained these data and their accompanying documentation on 11/28/2019 from https://aad.archives.gov/aad/popup-tech-info.jsp?s=5057

and the start and end position for each variable in the 60 fixed-width .txt files.

B Data Preprocessing

For each of the three types of records (applications, claims, and deaths), 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."

3. We dropped records without a valid birth date or records that had been anonymized by the Social Security administration, denoted by "ZZ."

B.1 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. Figure 6 gives a high-level overview of the process for constructing the BUNMD.

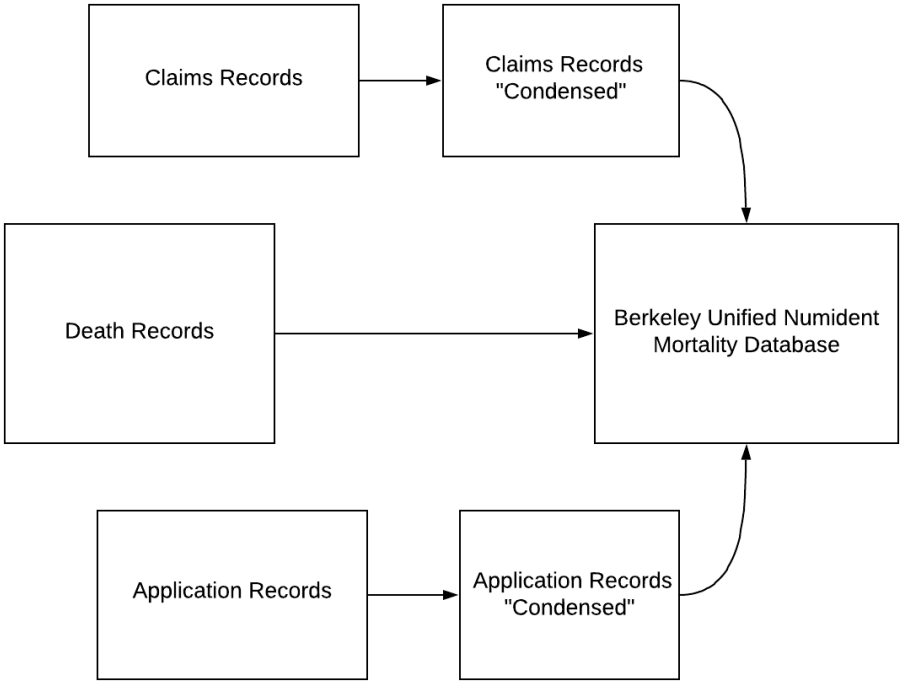

Figure 6: BUNMD creation flowchart. Applications and claims record with multiple entries were "condensed" into a single entry using the set of decision rules in Table 4.

While a person can only have one death entry, they might have several application or claim entries. Figure 7 shows the distribution of application and claim entries for those with a death record. 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. Therefore, 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 4). In order to study name changes, race changes, and other features, the original NARA Numident records are

useful and are available upon request.

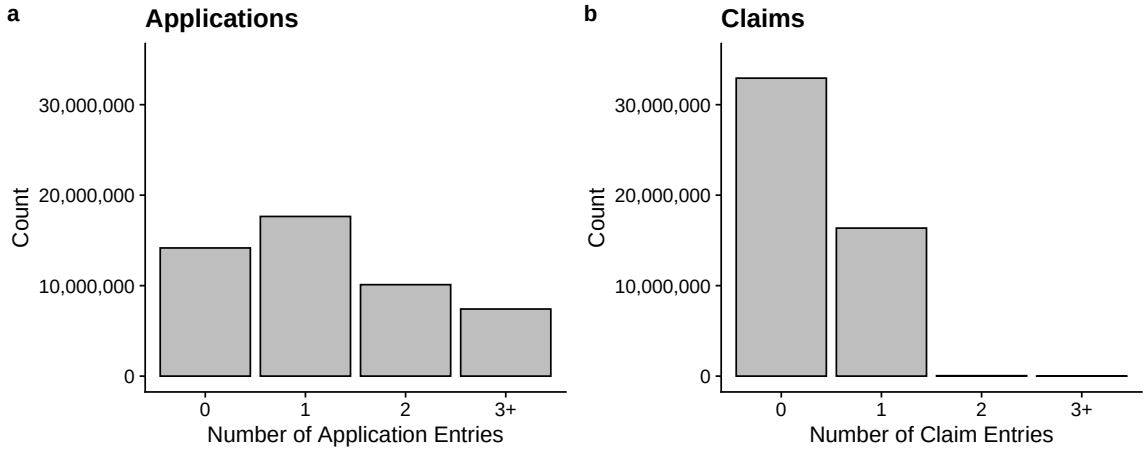

Figure 7: Number of entries per person for the Numident Application and Claim files.

B.2 Geographic variables

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 "socstate" variable reporting the state corresponding to the first three digits of the Social Security number using the crosswalk provided in Barron (1982). 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.

B.3 BUNMD Completeness

Most variables in the BUNMD are not available for every record. The "completeness" of each variable – the proportion of records with a non-missing value – varies across each variable. Figure 8 shows the completeness of each variable in the 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 4: The decision 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.

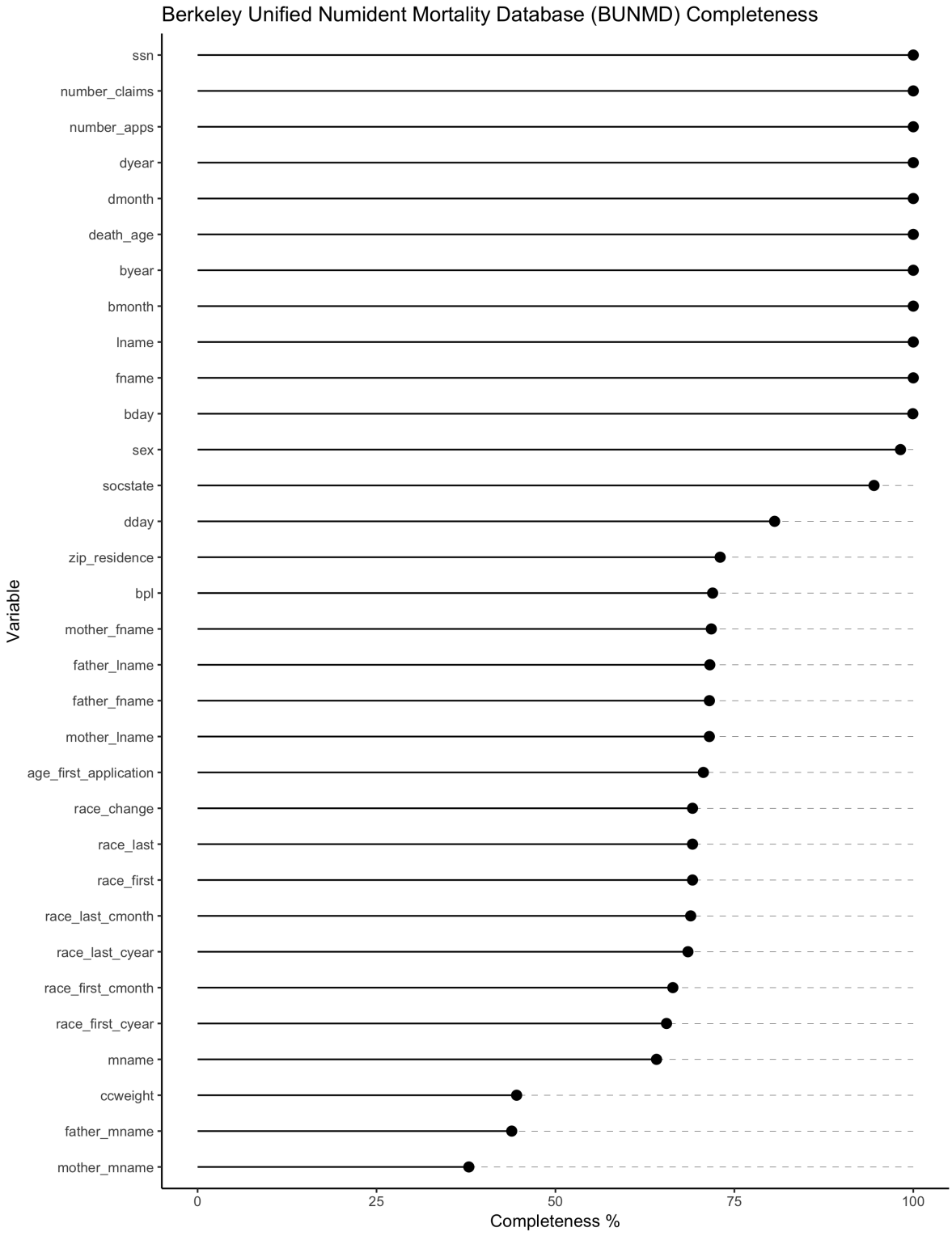

Figure 8: The completeness of each variable in the BUNMD.

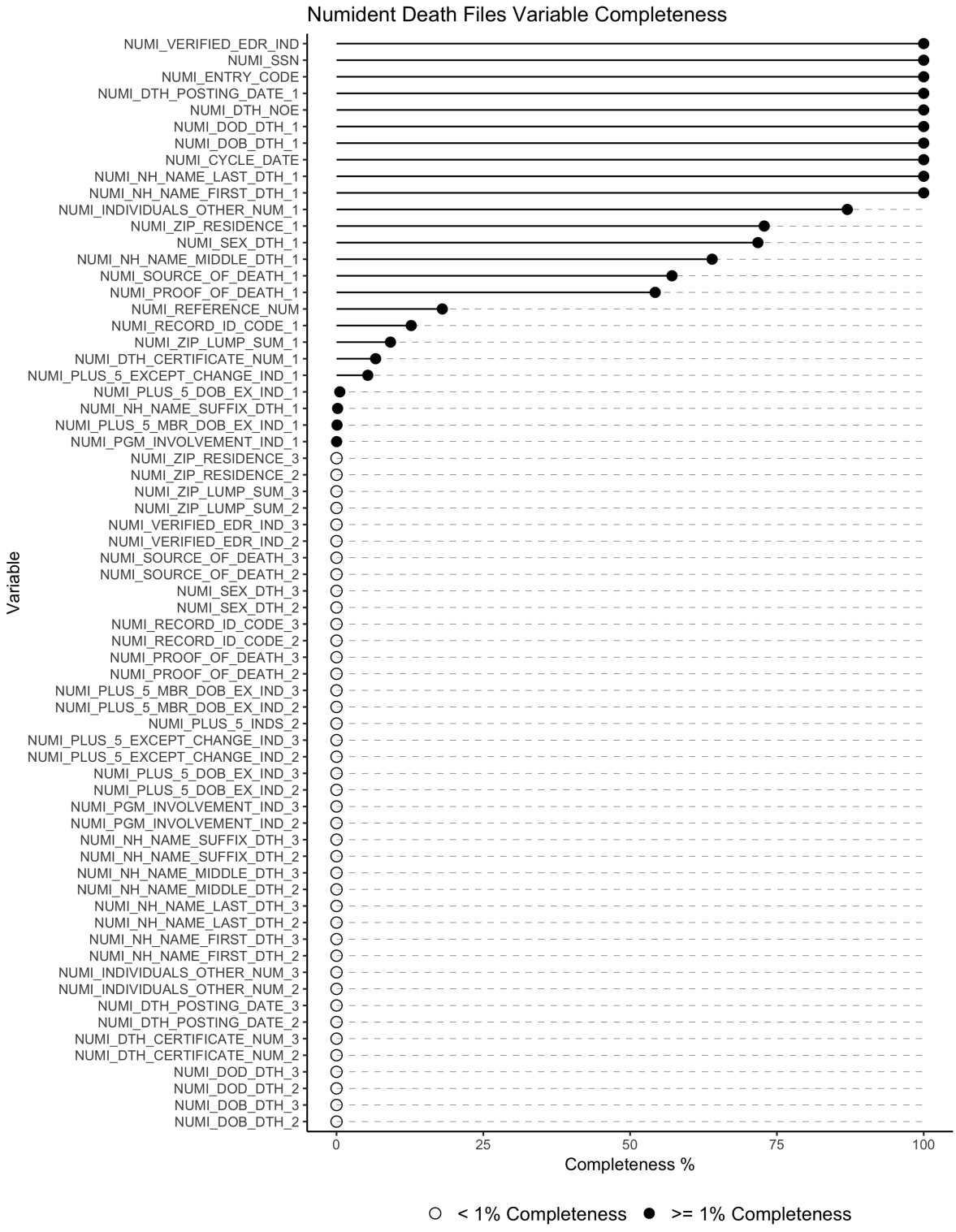

Figure 9: Completeness of the NARA Numident death records

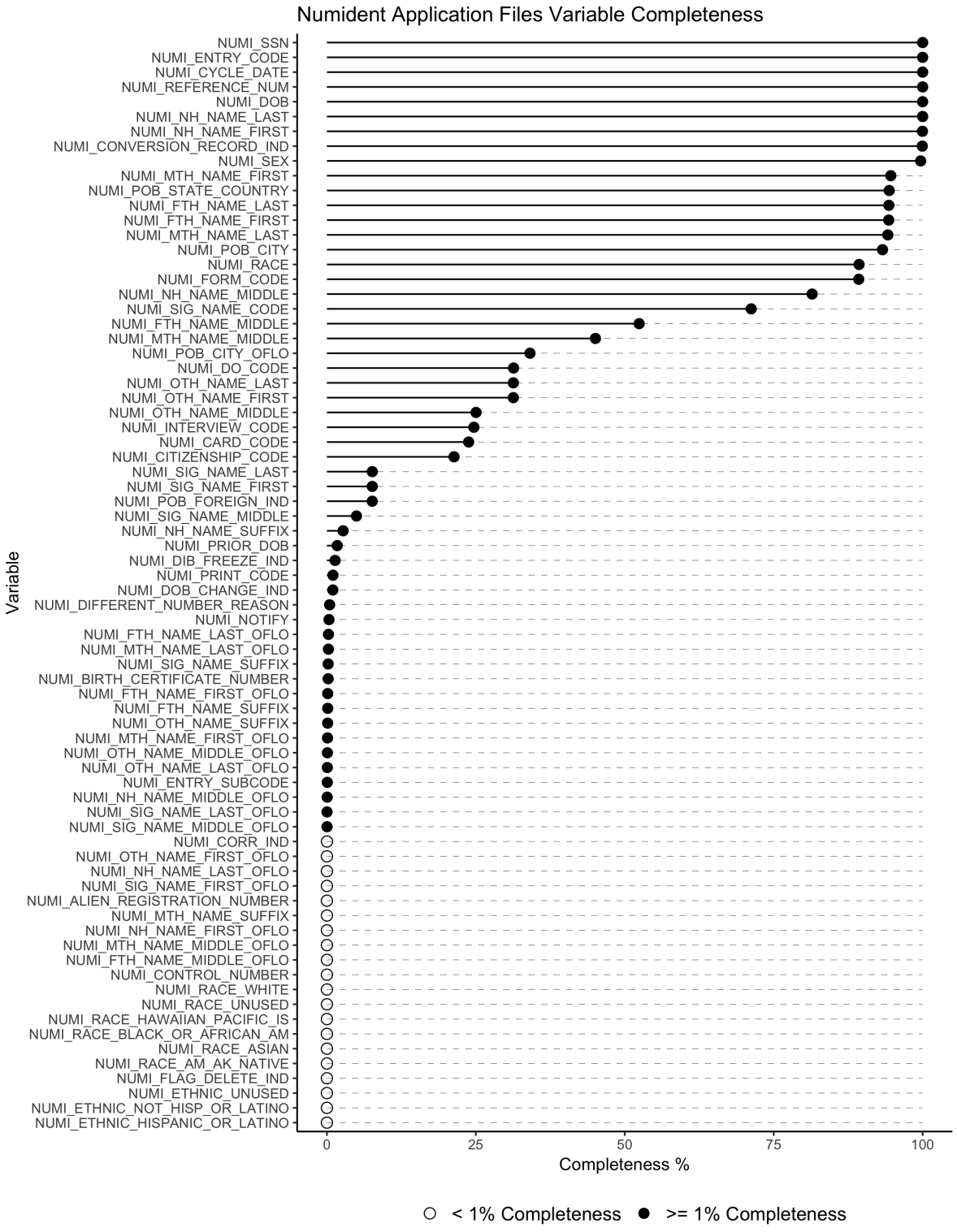

Figure 10: Completeness of the NARA Numident death records

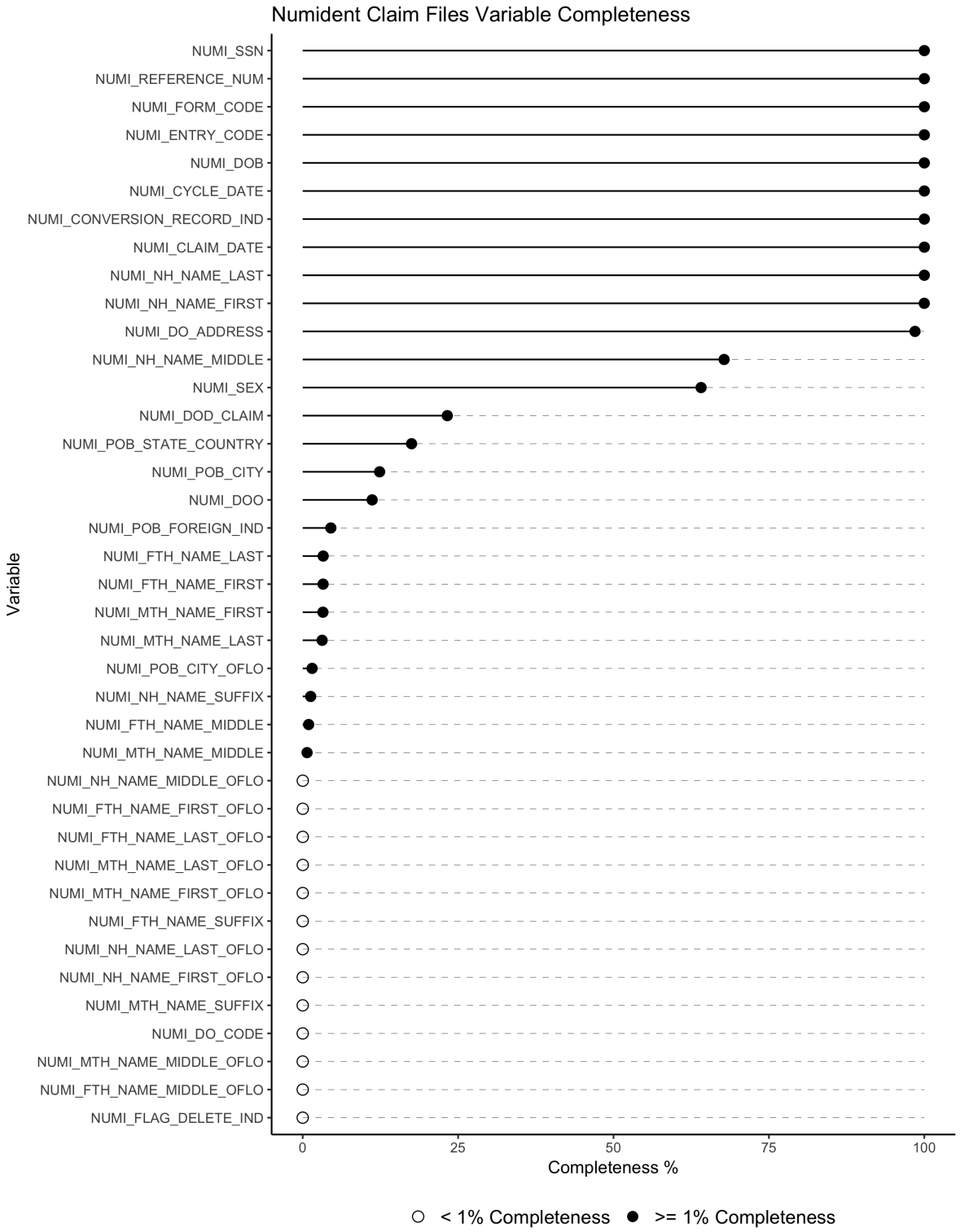

Figure 11: Completeness of the NARA Numident claims records

B.4 Replication Code

The full code and data to replicate the analyses in this paper are available on the following GitHub repository: https://github.com/caseybreen/bunmd. The code below replicates Figures 4 for men.

```
## Library Packages
library(data.table)
library(tidyverse)
library(gompertztrunc)

## Read in bunmd
bunmd <- fread(bunmd_v2.csv)

## Filter to "complete cases"
bunmd <- bunmd %>%
    filter(!is.na(ccweight)) %>%
    as.data.table()

## Create vector of countries for analysis
countries <- c("Canada", "Mexico", "Cuba", "Jamaica", "England", "Ireland", "Greece",
"Italy", "Portugal", "Austria", "Philippines", "Japan", "Syria", "Czechoslovakia",
"Poland", "Yugoslavia", "Russia", "China", "USA")

## Restrict to top countries
bunmd <- bunmd %>%
  filter(bpl_string %in% countries)

## Filter to birth cohorts and years of analysis
bunmd_analysis <- bunmd %>%
  filter(byear %in% 1910:1919 &
           dyear %in% 1988:2005 &
           age_first_application < 65)

## Set reference groups for model
bunmd_analysis <- bunmd_analysis %>%
  mutate(byear = relevel(as.factor(byear), ref = "1910"),
         country = relevel(as.factor(bpl_string), ref = "USA"))

## Restrict to men
bunmd_analysis_men <- bunmd_analysis %>%
  filter(sex == 1)

## Run Gompertz Parametric Model
gompertz_results <- gompertz_mle(death_age ~ country, data = bunmd_analysis_men,
                                 left_trunc = 1988, right_trunc = 2005,
                                 weights = ccweight)

## Convert hazards to life expectancy
gompertz_results_men <- convert_hazards_to_ex(gompertz_results$results, age = 65,
use_model_estimates = T)

## Plot Gompertz results for men
gompertz_results_men_plot <- gompertz_results_men %>%
  mutate(parameter = socviz::prefix_strip(parameter, prefixes = "country")) %>%
  ggplot(aes(y = reorder(parameter, e65), x = e65, xmin = e65_lower, xmax = e65_upper)) +
  geom_pointrange() +
  theme_cowplot() +
  labs(x = "e65_difference",
       y = "")
```