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

Joshua Goldstein[*]

Casey Breen[†]

20 January, 2020

Abstract

With the release of Social Security application and death records, the National Archives and Records Administration (NARA) has created a new administrative data resource for researchers studying mortality. This publicly available data provides demographic researchers access to over 49 million mortality records with rich geographic detail. We document the contents of this data set, provide free access to a cleaned and harmonized version of the data, and discuss statistical methods for estimating mortality differentials based on this deaths-only data set.

Introduction

The Numerical Identification System (Numident) forms the backbone of the U.S. Social Security Administration's record keeping system. For every person with a Social Security number, it tracks their earnings status, 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). NARA completed a public release of these records, offering nearly complete coverage of those who died from 1988 to 2005. In this paper, we describe the contents of the publicly available NARA Numident records, introduce a cleaned and harmonized version of the data, show how the records can be used for the study of mortality in the United States, and provide new methods for estimating mortality from death records.

---

[*]josh.goldstein@berkeley.edu; Department of Demography, UC Berkeley

[†]caseybreen@berkeley.edu; Department of Demography, UC Berkeley

The NARA Numident is an individual-level mortality data set with over 49 million death records. It includes variables describing race, sex, birthplace, ZIP Code of residence at time of death, and administrative variables, such as a person's age on their first first Social Security application and total number of Social Security applications. Notably, the death coverage is nearly complete for deaths to persons age 65+ between 1988-2005.

As death coverage is only high for a small window, mortality estimation with the NARA Numident records presents challenges. Survivorship cannot be estimated, as the deaths do not have denominators. Further, the observed deaths are left and right truncated, which makes calculating unbiased estimates of mortality difficult. We introduce methods to overcome this. Another drawback of the Numident records is that they lack any indicator of socioeconomic status. To overcome this, the records can be linked to other data sources, either by Social Security number or a combination of other key identifiers.

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

Another publicly available data resource for mortality research is the Social Security Death Master File (DMF). The DMF was first made available in 1988, and is extracted quarterly from the Numident records (Hill and Rosenwaike, n.d.). The file has been used by some researchers to study mortality, particularly at older ages (Gavrilov and Gavrilova 2012). While the file has high death coverage for deaths of persons aged 65 or older from 1975 - 2005, it lacks the covariates, such as sex, race, or place of birth, available in the NARA Numident records (Hill and Rosenwaike, n.d.). The methods we provide here are also useful for researchers working with the DMF.

We are also in the process of linking both the DMF and the NARA Numident records to the full-count 1940 Census, to create a rich, public linked administrative dataset for the study of mortality (CENSOC).
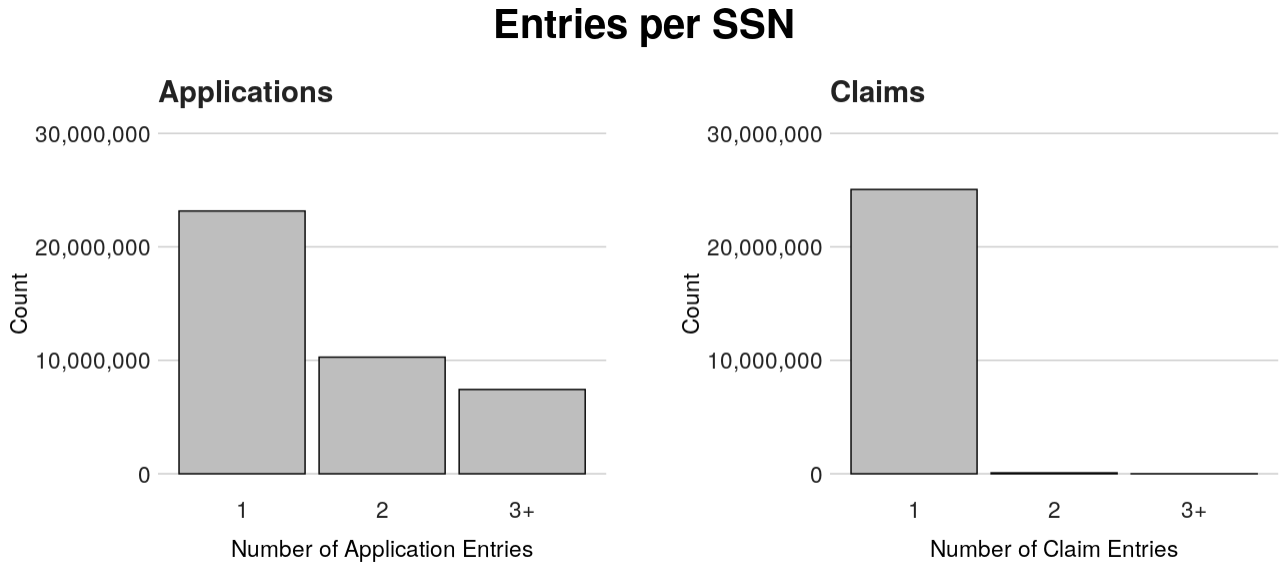
Figure 1: Number of entries per person for the Numident Application and Claims files.

The Content of the NARA Numident files

The NARA Numident records contain three types of entries: applications, claims, and deaths. The Social Security Administration adds a new entry to the Numident when a Social Security cardholder submits a new application or claim. New entries never overlay old entries. Instead, a new entry is added to the Numident. This ensures that no information is never overwritten in the Numident. Figure 1 shows the distribution of application and claim entries per person. While 43.3% of persons in the Numident have multiple application entries, only 0.3% of persons have multiple claim entries. No persons have multiple death entries.

To illustrate the structure and content of the publicly-released National Archive Numident files, we show in figure 1 the records that were released for the actress Lana Turner who died in 1995 and for the Supreme Court Justice Thurgood Marshall, who died in 1993.

For Thurgood Marshall, we have one application and one death record:

Application Record:

ssn	citizenship_code	cycle_date	entry_code	dob	sex
131074264		193712XX	0	07021908	1

race	pob_state_country	year_cycle	month_cycle	pob_foreign_ind
2	MD	1937	12	

fname	mname	lname	mother_fname	mother_mname	mother_lname
THURGOOD		MARSHALL	NORMA	A	WILLIAMS

father_fname	father_mname	father_lname
WILLIAM	C	MARSHALL

Death Record:

ssn	sex	zip_residence	lname	mname	fname	byear
131074264	1	220411335	MARSHALL		THURGOOD	1908

dyear	socstate	bmonth	dmonth	bday	dday
1993	36	7	1	2	24

Table 1: NARA Numident Application and Death Record for Supreme Court Justice Thurgood Marshall.

For Lana Turner we have four application records, corresponding it seems to name changes each time she was married.

Application Record:

ssn	citizenship_code	cycle_date	entry_code	dob	sex	race	pob_state_country
567183907		193703XX	0	02081921	2	1	ID
567183907		194907XX	2	02081921	2	1	ID
567183907		195611XX	2	02081921	2	1	
567183907		197009XX	2	02081921	2	0	UN

pob_foreign_ind	fname	mname	lname	mother_fname	mother_mname	mother_lname
	LANA		TURNER	MILDRED	F	COWAN
	LANA	TURNER	TOPPING	MILDRED	F	TURNER
	LANA	TURNER	BARKER			
	LANA	TURNER	DANTE			

father_fname	father_mname	father_lname	year_cycle	month_cycle
JOHN	M	TURNER	1937	3
JOHN	V	TURNER	1949	7
			1956	11
			1970	9

Death Record:

ssn	sex	zip_residence	lname	mname	fname	byear	dyear	socstate	bmonth
567183907	2	900255240	TURNER		LANA	1921	1995	6	2

dmonth	bday	dday
6	8	29

Table 2: NARA Numident Application and Death Record for actress Lana Turner.

In order to facilitate analysis of the NARA Numident records, we have created the Berkeley Unified Numident Mortality Database. This file condenses the Numident death, application, and claims records into a single file with one record per person. This file is available for download at _____ . The file includes about 49 million records, 28 variables, and is about 5.7 Gb in size.

For Lana Turner, the BUNMD data record is:

ssn	zip_residence	lname	mname	fname	byear	dyear	socstate	bmonth	dmonth
567183907	900255240	TURNER		LANA	1921	1995	600	2	6

bday	dday	number_apps	race	race_change	number_claims	sex	bpl	father_fname
8	29	4	1	0	0	2	1600	JOHN

father_mname	father_lname	weight	death_age	cweight
NA	TURNER	1.047614	74	1.105444

To facilitate analysis of the NARA Numident records, we introduce a cleaned and harmonized version of the data— the Berkeley Unified Mortality Numident Database (BUNMD). The original NARA Numident release contained 49,459,293 death records entries, 72,120,516 applications entries corresponding to 40,870,455 unique individuals, and 25,228,257 claims entries corresponding to 25,140,847 unique individuals. The BUNMD condenses those records into one file with a single record per person. The BUNMD contains every death entry in the NARA Numident. We then use a set of decision rules to select a single value of birthplace, race, sex, etc. across all an individual's application and claims entries. We construct variables reporting (1) total number of application, (2) total number claim entries, (3) age of first application, (4) state the Social Security number was issued. Figure 2 shows the process for constructing the BUNMD. Please see the technical appendix for more details on the construction of this file. In order to study name changes, race changes, and other features, the original NARA Numident records are useful. They are available upon request.
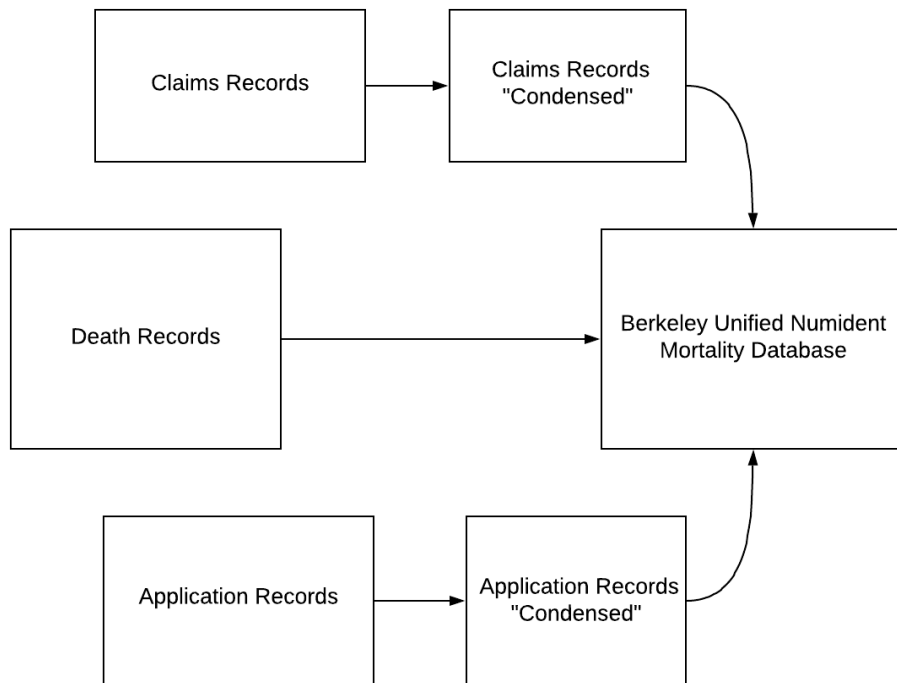
Figure 2: Berkeley Unified Numident Mortality Database creation flowchart.

Numident Coverage

The NARA Numident records are a subset of the master Numident records. NARA documentation states that their released records are for individuals with a verified death between 1936 and 2007 or persons over 110 years of age as of December 31st, 2018. This alone does not explain the coverage patterns we observe. Figure 3 compares the total number of deaths for persons age 65+ in the BUNMD to the Human Mortality Database (HMD). Death Coverage is nearly complete between 1988 and 2005. Figure 4 shows the coverage visualized on a age-period Lexis surface, an established demographic visualization technique. The left panel shows the BUNMD death coverage — the proportion of total deaths in the United States captured in the BUNMD — for ages 0 to 100 from 1940 to 2010. Coverage is low. The right panel zooms in on the relatively high coverage death coverage for persons dying between age 65 and 100 between 1985 and 2005. This is an important limitation of the BUNMD dataset.

To address this, we create two BUNMD samples to allow researchers to use mortality records when coverage is most complete . The Sample 1 includes deaths to persons age 65+, occurring between 1988 and 2005, from the cohorts of 1900-1940. Sample 2 is the subset of Sample 1 records with complete information on sex, birthplace, and race. For each sample,

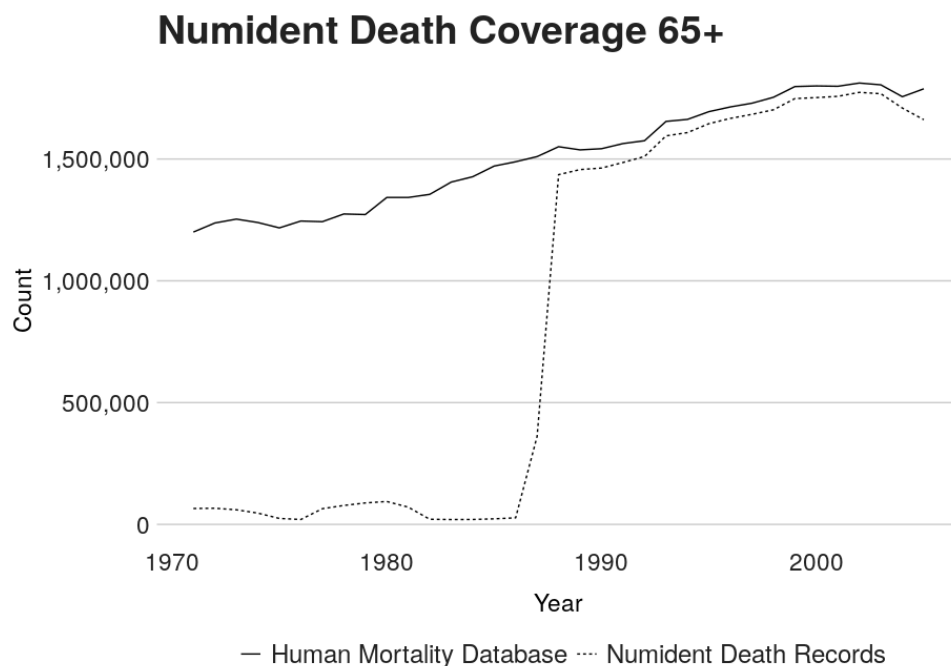## Numident Death Coverage 65+

Figure 3: BUNMD Death Coverage for persons 65+

we constructed inverse probability weights to the Human Mortality database on age at death, year of birth, year of death, and sex.

Notes: - the material below we've mostly done — we're going to weight by age$dyear$sex. TODO - check on race - check on state - race (maybe NCHS) see Berkeley Mortality Database for nice versions of Black, White death counts. - key thing is if inverse HMD weights which are not based on race work. - real issue here is weights and their pros and cons. and what we can give evidence for about pros and cons. - Variables

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

8

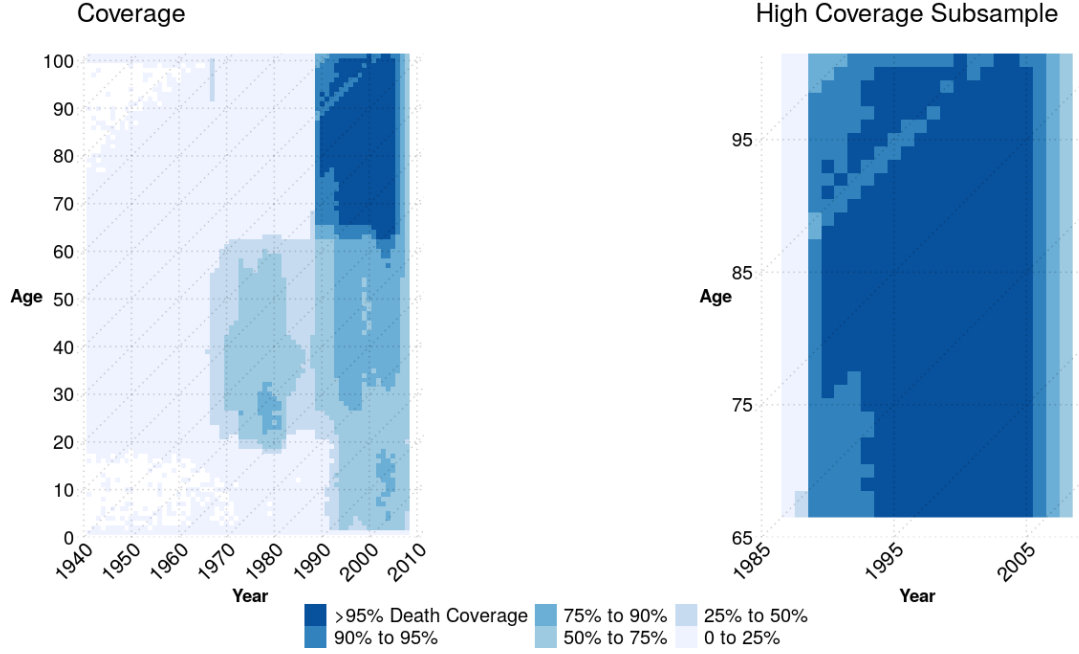## Numident Death Record Coverage

Figure 4: Lexis diagram of BUNMD death coverage

regression, inflating the observed coefficients in order to account for truncation. Finally, we introduce the cox regression method

Method 1: Parametric survival models

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

$$h(x) = a * exp(b * x) \tag{1}$$

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

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

Finally, at older ages, although there is still much debate, there may be a leveling of mortality.

9

Thus the logistic ... model has been introduced to account for this leveling of mortality. For any parametric model, it is possible to write down the likelihood given the deaths we observe. For truncated cohorts, with known left-truncation L and known right truncation R, we can write down the likelihood as

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

The estimates of the vector θ of parameters can be obtained by maximizing the likelihood, or, equivalently, the log-likelihood. We have written functions in the computer language R that can be used to obtain maximum likelihood estimates of these parametric models.

Method 2: Ordinary Least Squares Regression

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

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

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

Truncation, however, will tend to bias downward the estimated effects of any covariates. (need citation). Truncaction excludes the tails of the distribution, thus reducing the average difference between groups.

The idea is that the average differences between groups will be measured to be much smaller if we exclude the tails of the distribution.

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

Method 3: Cox regression for extinct cohorts

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

Case Studies

Geography

There are several geography variables in the BUNMD. The application entries have information on birthplace. For persons born in the United States, the geographic resolution is state-level. For persons born outside of the United States, the geographic resolution is country-level. The Numident death entries contains the 5 or 9-digit ZIP code of the residence at the time of death. ZIP Codes as a unit for spatial analysis can offer insights into a variety of spatial questions, but have an accompanying set of challenges ((**???**)).

Figure 5 shows life expectancy at age 65 by ZIP Code in Ohio for the pooled for the cohorts of 1915 to 1920. Figure 6 focuses on Cleveland's Cayuga County. Life expectancy is lower in Cleveland, and higher in the surrounding suburbs. These old-age mortality disparities — well-documented in the United States — are likely driven by racial segregation.

Gompertz Mortality

The BUNMD can also be used to looks at race differentials in mortality by state. One can look at how disparities in old-age mortaltiy changes over time within a given state, and look at within cohort black/white mortality crossovers. In Figure 7, we compare estimates of life expectancy at age 65 for Whites and Blacks change for the cohorts of 1900 to 1920 for the state of Alabama.

Conclusion

The NARA Numident release has created a new administrative data resource for researchers studying mortality. We introduce a publicly-available cleaned and harmonized version of the dataset with over 49-million deaths. The high spatial resolution allows for fine-scaled geographic studies, and demographic covariates. We introduce methods to overcome a
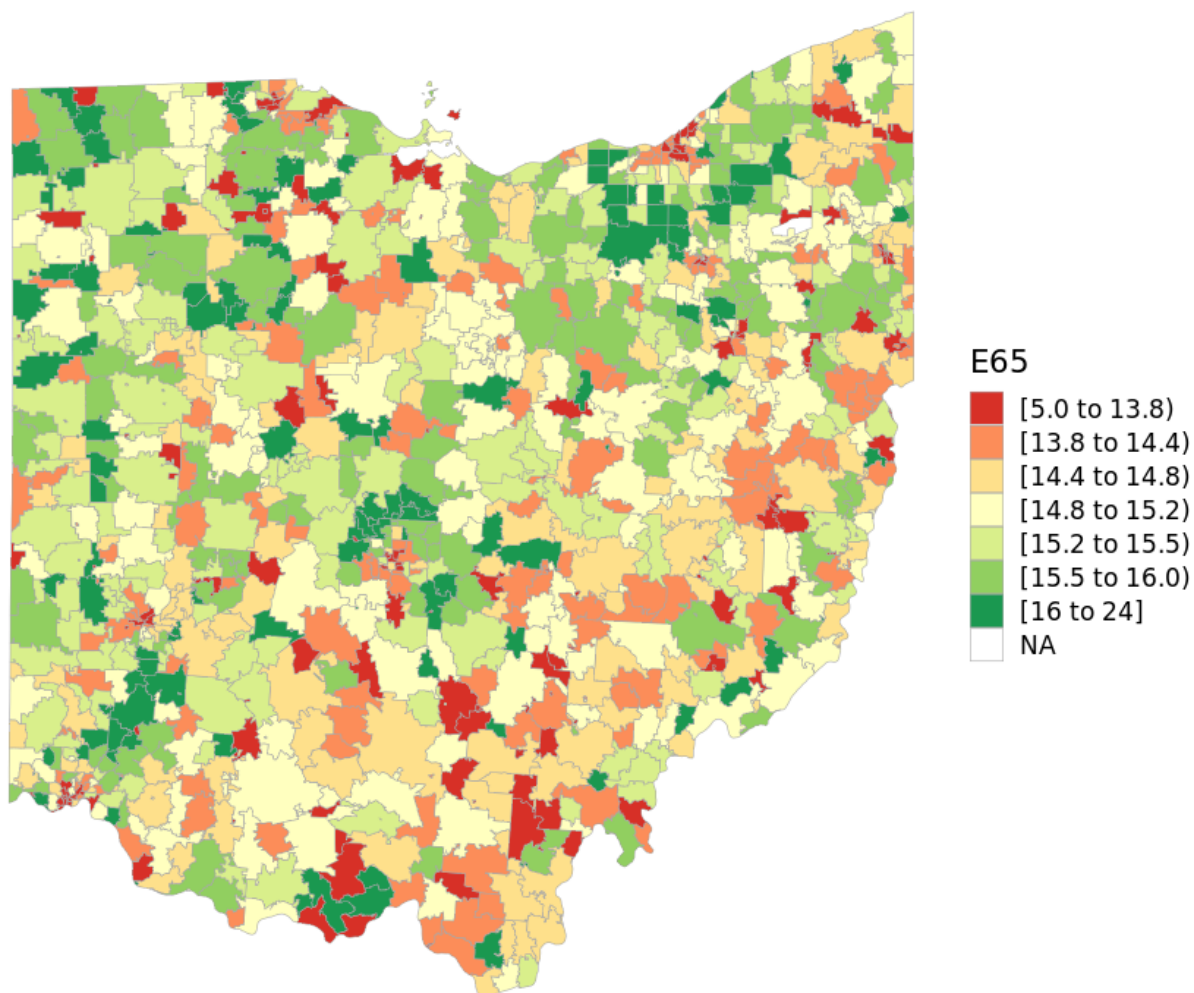
Life Expectancy at age 65: Ohio

Figure 5: Ohio life expectancy at age 65 by ZIP Code for pooled cohorts of 1915 to 1920.

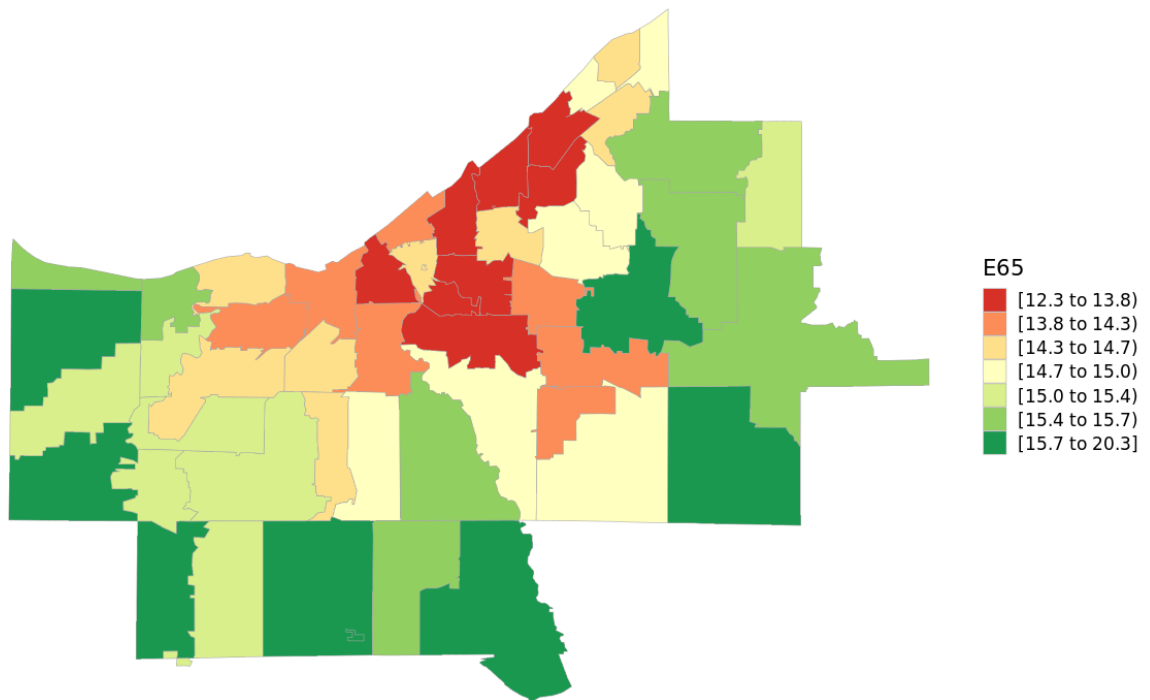## Cuyagoa County (Cleveland) Life Expectancy at Age 65

Figure 6: Life expectancy at age 65 by ZIP Code for pooled cohorts of 1915 to 1920.

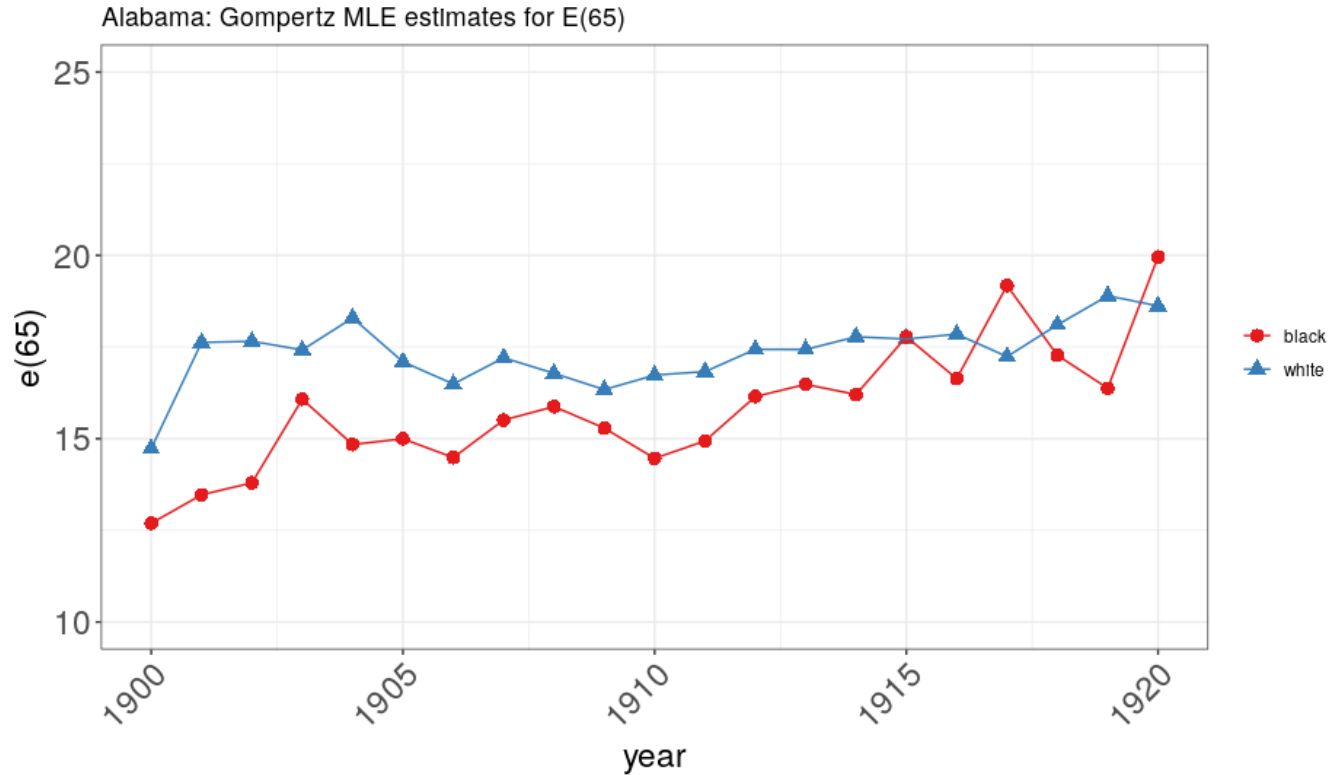shortcoming of the dataset, namely that the death coverage is only high for a narrow window of deaths.

Figure 7: Gompertz E65 estimates for Alabama for Whites and Blacks.

Notes

- can get old-age mortality using reverse survival (good application would be black/white , elo, preston et al)

? other approaches . . . no, but can mention in a few sentences and refer to Monica's paper / dissertation

Case studies:

Pick a subset that are the BEST. - Race x state (just black and white) - BPL x time of 1st application (proxy for time of arrival – Mexicans and Fillipinos) - Something with own name and name of parents - [So what this means is that we want "race", "bpl", time of first application, anything we might have on citizenship (but that's not crucial for the 1st data set production), and then we want all the name info there is]

Coverage

- state, not really for substance, but still good for coverage

- first name, last name

- could do an ancestry corrector. my guess is that last name doesn't matter much. and neither does 1st name . . . but i don't know.

- Nicknames — as a proxy for social class. can distinguish blacks and whites.

- Yes, by country . . . super interesting but a lot of detail (e.g., when people came to the U.S. that is a bit hard to infer — but can use 1st date . . . this is kind of a paper in and of itself)

- n_ss5 applications: maybe a proxy for carelessness losing cards, life instability

- daylight savings time . . .? No — too hard a topic . . . but not bad for another paper

- ZIP code (say by income or house value or average educ): one illustrative example, e.g., Manhattan or NYC or Ohio

- Choose a case study that data is available for (like from NCHS) to validate our approach (daylight savings time?); (state?)

- Choose case studies that can't be done with other data . . . (race x daylight savings time?)(BPL !!!)(anything involving names) (anything involving exact date of birth).

- Influenza?????

- Weighting and HMD comparison (casey can do this with Josh's old code – but it does need to be adapted, because I was using pre-tabled data from HMD, and if you use numideath, you have to table it yourself into a Dx vector with ages of death as names(Dx) for the functions to work) (also find an example where weighting makes a difference)

- Truncated gompertz vs regression (but this can be in one of the case studies)

- Case studies: all on age at death

- Race x state

- Nicknames (use 1940 to get social class score of nicknames . . .?)

Public distribution, acknowledgement, conditions

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

For: Demographic Research or Demography (DR has the advantage of infinite extra cool links etc and being totally free and accessible — which is probably what we want; prestige of Demography not really needed for this paper (and it might not get accepted because of its nature).

References:

Black, Dan. n.d. "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.

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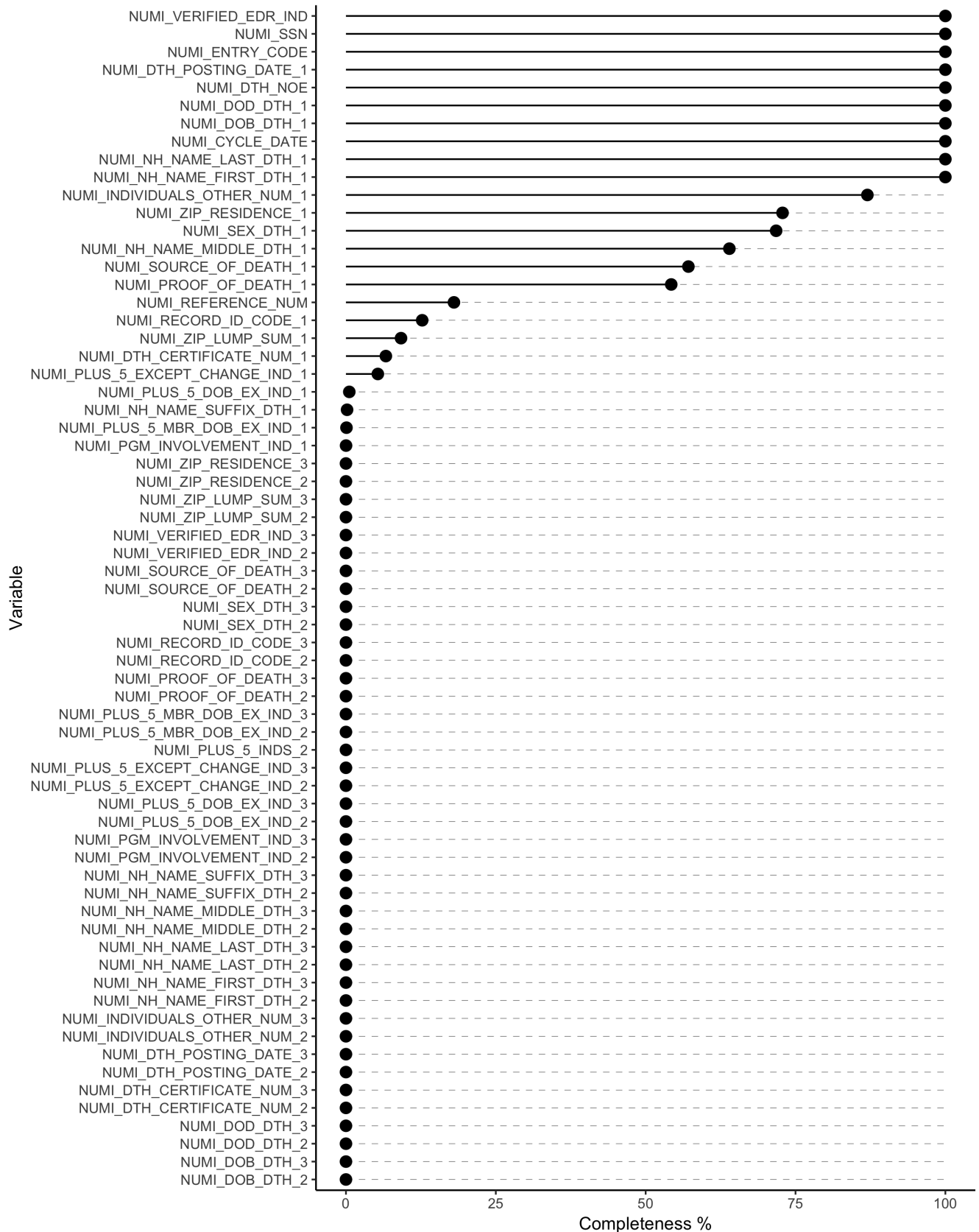
Puckett, Carolyn. 2009. "The Story of the Social Security Number." *Social Security Bulletin* 69 (2): 21.

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

Berkeley Unified Numident Mortality Database Completeness
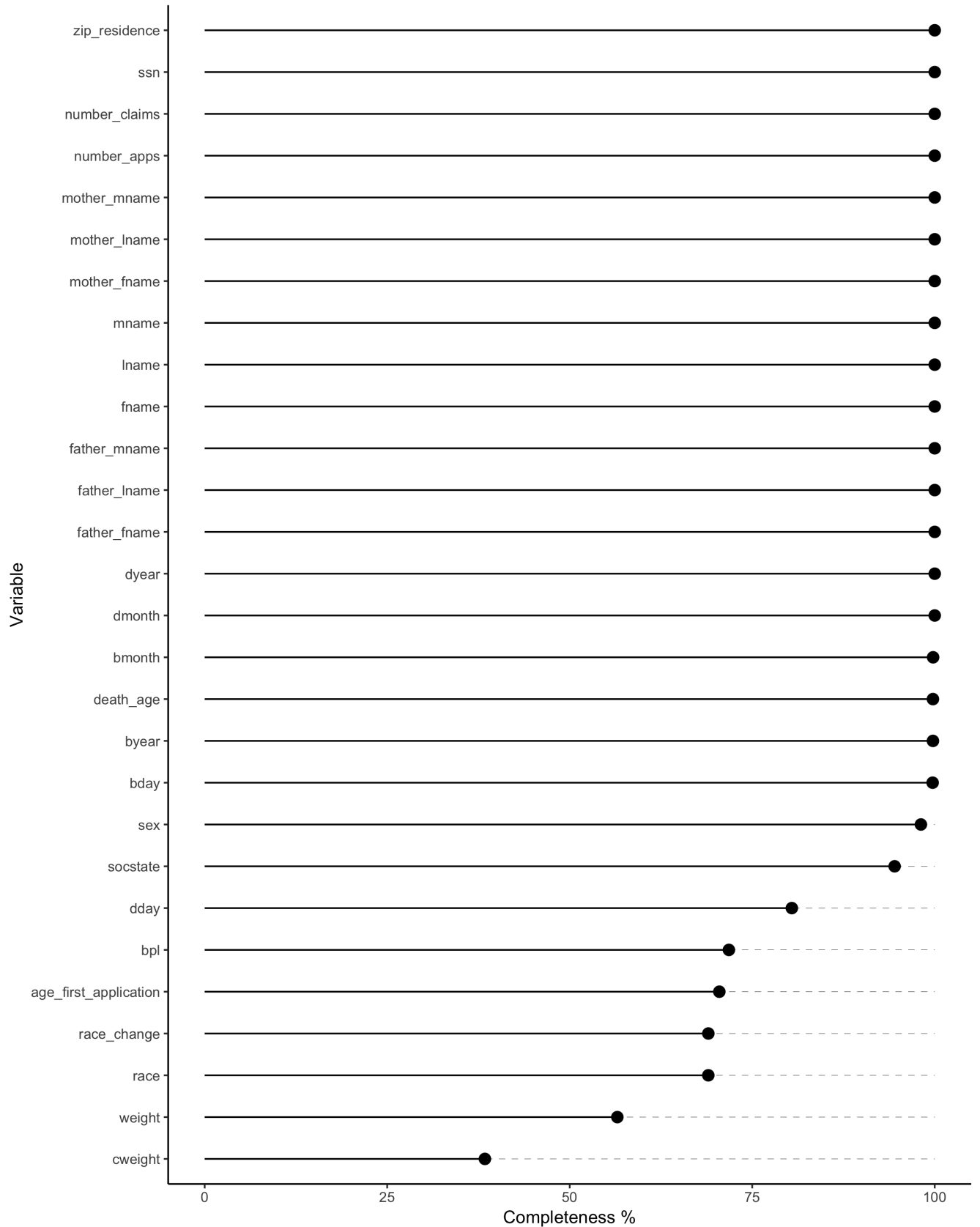
Numident Death Files Variable Completeness

Numident Application Files Variable Completeness

Numident Claim Files Variable Completeness