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

Joshua Goldstein*
Casey Breen†

21 February, 2020

Abstract

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

Introduction

The Numerical Identification System (Numident) forms the backbone of the U.S. Social Security Administration's record keeping system. For every person with a Social Security number, the Numident tracks 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 publicly released these records in 2019, offering nearly complete coverage of those who died from 1988 to 2005. In this paper, we describe the contents of the publicly available

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

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.

The NARA Numident is micro-level dataset with over 49 million death records. It 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 at the individual level to other data sources, either by Social Security number or a combination of other key identifiers, to obtain income and education covariates (Goldstein et al. 2019). The death coverage is nearly complete for deaths to persons age 65+ for the window of 1988-2005. The large size of the NARA Numident will open up new avenues for research into smaller population subgroups and disparities later in the life-course, when cohorts are a fraction of their original size.

There are several specific considerations for using the NARA Numident records for mortality estimation. As the mortality records only include deaths, there is no measure of survivorship. Mortality rates must be estimated without denominators, so traditional statistical methods relying on exposure to risk are not appropriate. Additionally, the observed deaths are left and right truncated, which makes calculating unbiased estimates of mortality difficult. We introduce appropriate methods for estimating mortality with truncated data in the absence of accurate exposure to risk measures.

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

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

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

The Content of the NARA Numident files

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

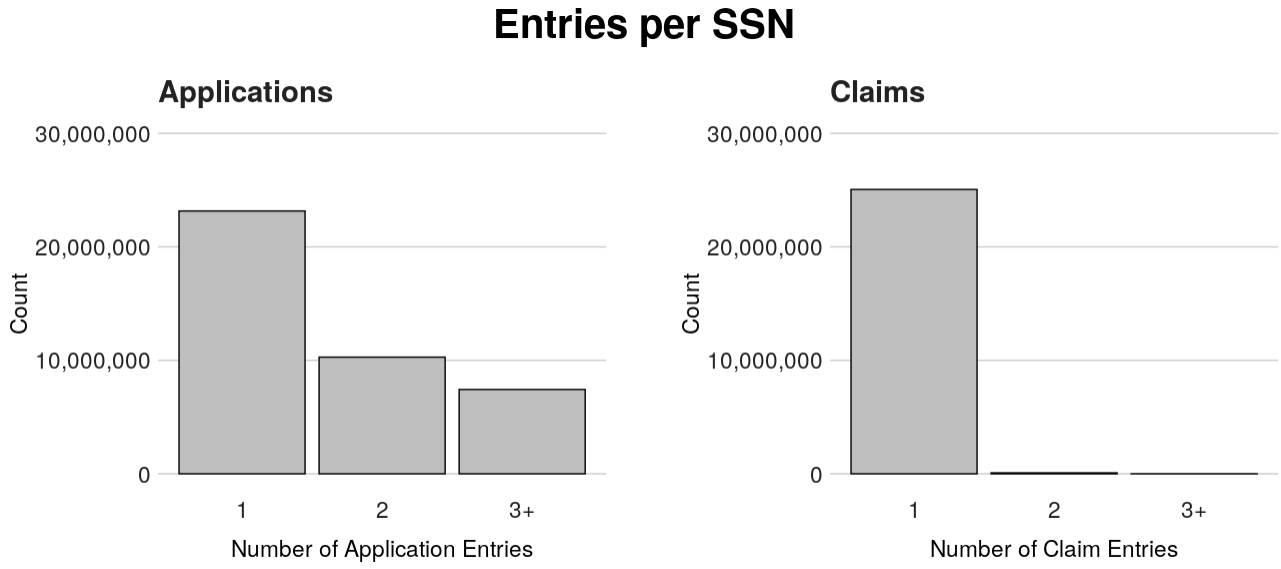


Figure 1: Number of entries per person for the Numident Application and Claims files.

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

Thurgood Marshall

NARA Numident Record

entry	ssn	cycle_date	dob	sex	race	pob	fname	mname	lname	mother_fname	mother_lname	father_fname	father_lname
application 1	131074264	12/1937	07/02/1908	1	2	MD	THURGOOD		MARSHALL	NORMA	WILLIAMS	WILLIAM	MARSHALL

entry	ssn	sex	zip_residence	lname	mname	fname	byear	dyear	socstate	bmonth	dmonth	bday	dday
death	131074264	1	220411335	MARSHALL		THURGOOD	1908	1993	36	7	1	2	24

Berkeley Unified Numident Mortality Database (BUNMD) Record

ssn	zip_residence	lname	fname	byear	dyear	socstate	bmonth	dmonth	bday	dday	race	race_change	sex	bpl	death_age
131074264	220411335	MARSHALL	THURGOOD	1908	1993	36	7	1	2	24	2	0	1	2400	84

Lana Turner

NARA Numident Record

entry	ssn	cycle_date	dob	sex	race	pob	fname	mname	lname	mother_fname	mother_lname	father_fname	father_lname
application 1	567183907	03/1937	02/08/1921	2	1	ID	LANA		TURNER	MILDRED	COWAN	JOHN	TURNER
application 2	567183907	07/1949	02/08/1921	2	1	ID	LANA	TURNER	TOPPING	MILDRED	TURNER	JOHN	TURNER
application 3	567183907	11/1956	02/08/1921	2	1		LANA	TURNER	BARKER				
application 4	567183907	09/1970	02/08/1921	2	0	NA	LANA	TURNER	DANTE				

entry	ssn	sex	zip_residence	lname	mname	fname	byear	dyear	socstate	bmonth	dmonth	bday	dday
death	567183907	2	900255240	TURNER		LANA	1921	1995	6	2	6	8	29

Berkeley Unified Numident Mortality Database (BUNMD) Record

ssn	zip_residence	lname	fname	byear	dyear	socstate	bmonth	dmonth	bday	dday	race	race_change	sex	bpl	death_age
567183907	900255240	TURNER	LANA	1921	1995	6	2	6	8	29	1	0	2	1600	74

Table 1: A stylized example of how the NARA Numident records were combined to create the BUNMD. For this specific example, several variables were excluded (e.g. person's middle names, parent's names, etc.).

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

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

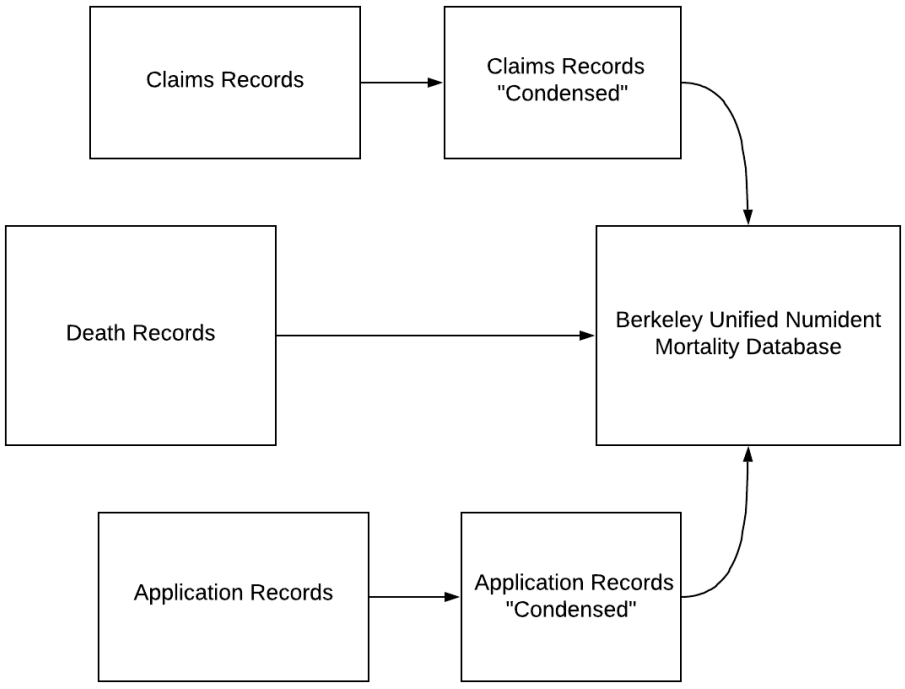


Figure 2: Berkeley Unified Numident Mortality Database creation flowchart.

Numident Coverage

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

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

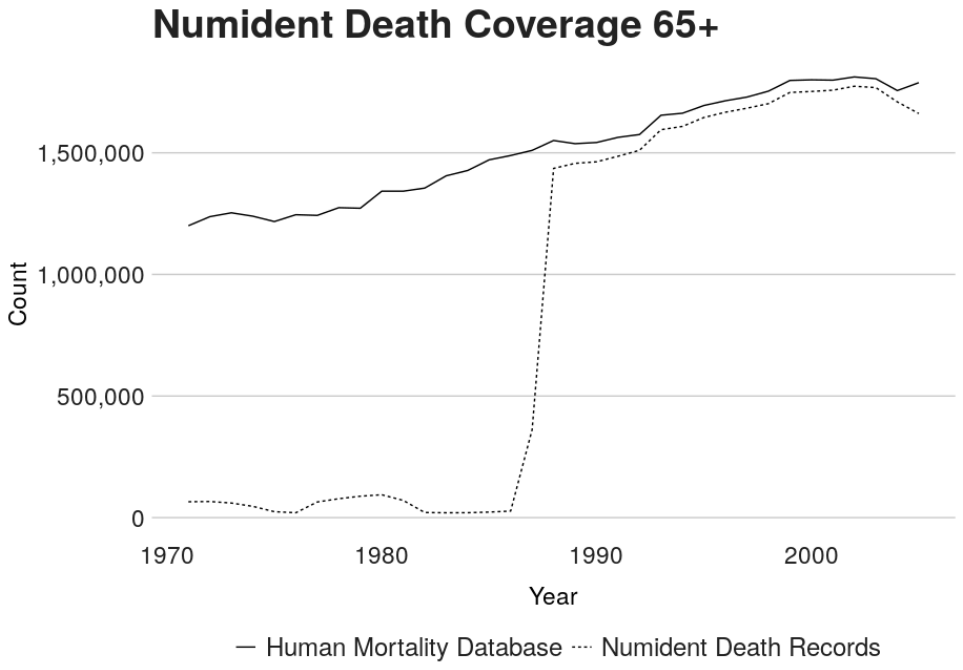


Figure 3: BUNMD Death Coverage for persons 65+

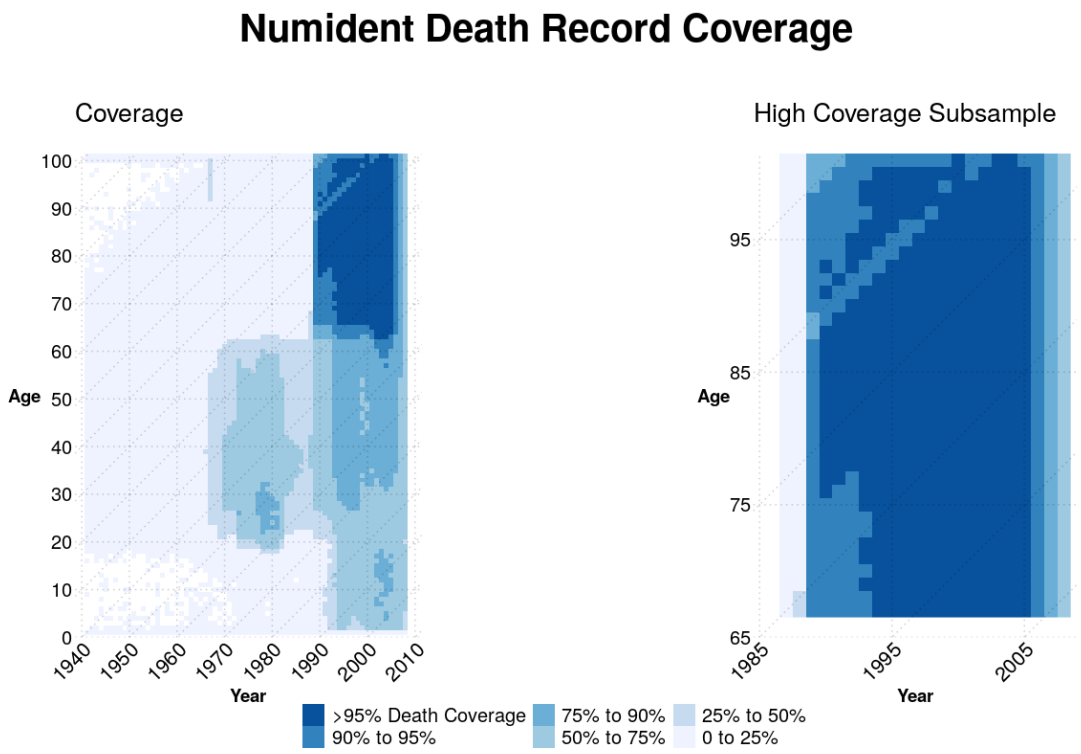


Figure 4: Lexis diagrams of BUNMD death coverage. The left panel shows the BUNMD death coverage for 1940 to 2010. The right panel zooms in on age-period area where coverage is highest.

Estimation: Deaths without Denominators

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

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

Method 1: Parametric survival models

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

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

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

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

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

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

The estimates of the vector θ of parameters can be obtained by maximizing the likelihood,

or, equivalently, the log-likelihood.

Method 2: Ordinary Least Squares Regression

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

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

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

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

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

Method 3: Cox regression for extinct cohorts

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

Case Studies

Geography

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

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

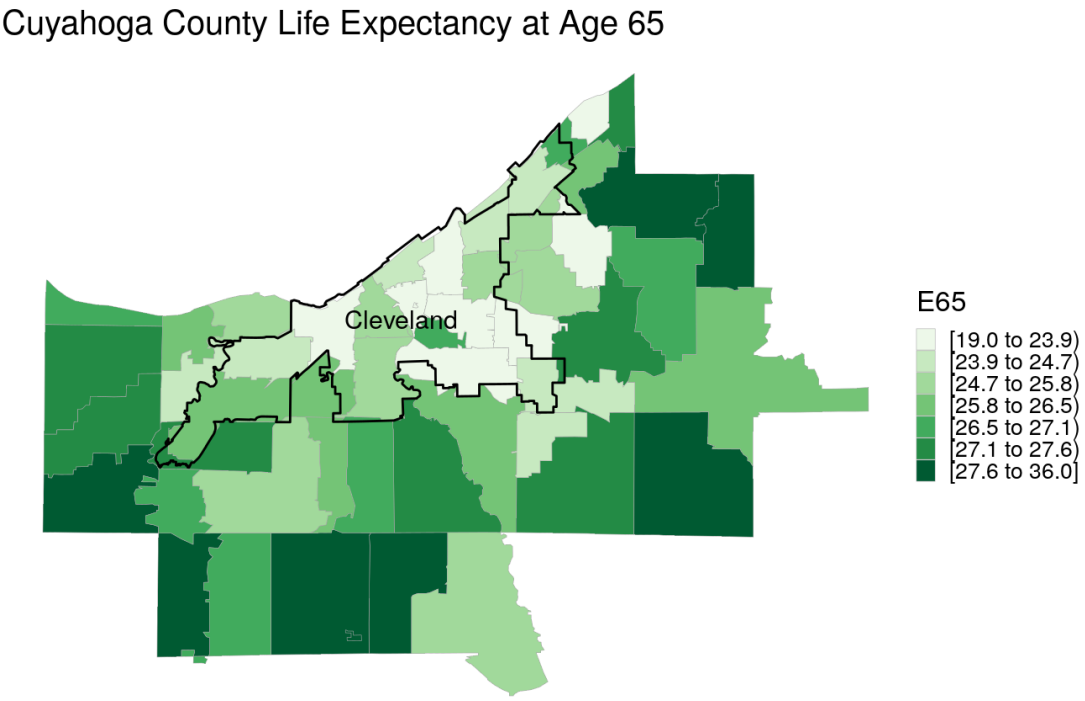


Figure 5: Life expectancy at age 65 in Cuyahoga County for the birth cohort of 1900.

Gompertz maximum likelihood estimation: race differentials

Gompertz maximum likelihood estimation can also be used to look at race differentials in mortality by state in the BUNMD. This method combines the observed distribution

of deaths over a certain window of deaths with external knowledge of human mortality age-patterns, allowing us to estimate mortality rates given a truncated window of deaths. We are assuming that the Gompertz model is appropriate and that the deaths we observe reflects the true population cohort distribution. Under-c verage will not bias estimates as long as the under-coverage is happening at random.

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

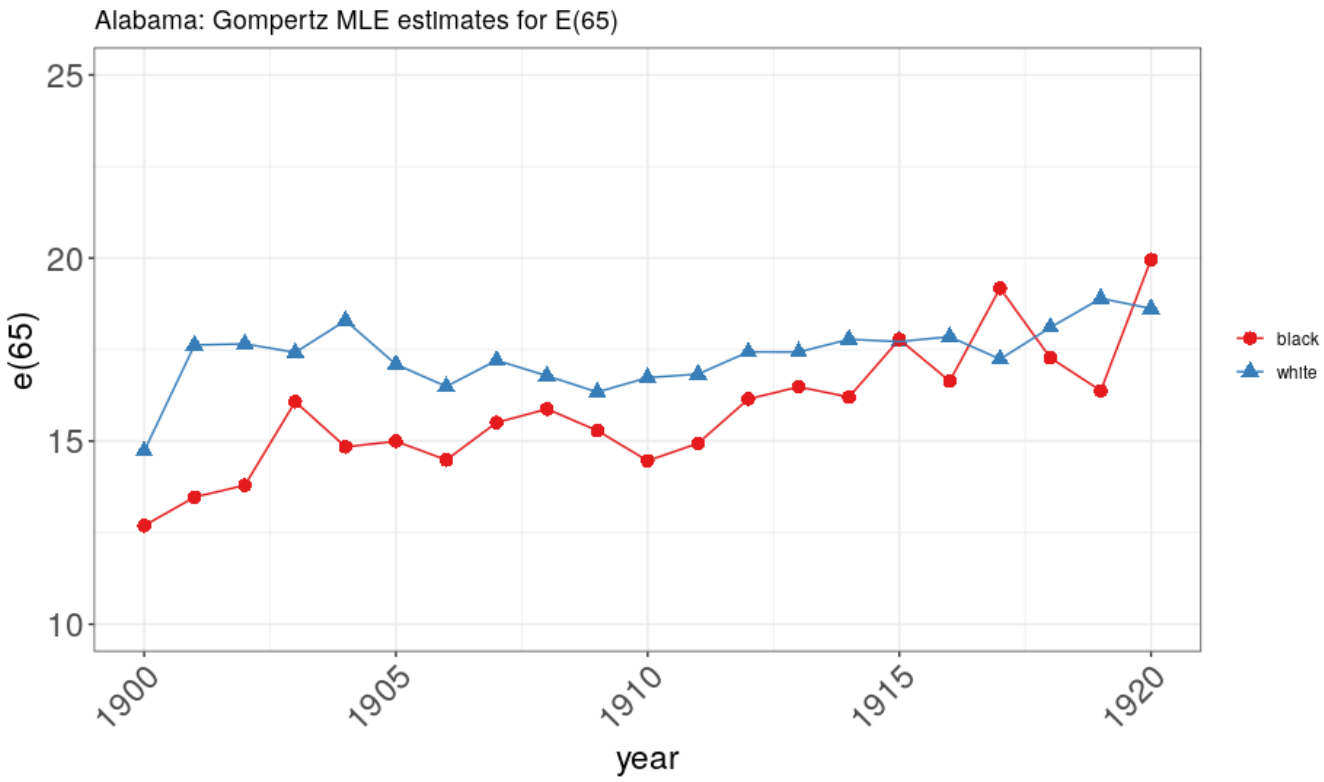


Figure 6: Gompertz E65 estimates for Alabama for Whites and Blacks.

Conclusion

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

the open-access nature of the data ensures that research is reproducible and extendable.

Public distribution, acknowledgement, conditions

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

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

References

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

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.

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.

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

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

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

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

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

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

Technical Appendix

This supplementary appendix presents the procedure for creating the Berkeley Unified Numident Mortality Database (BUNMD).

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, along with 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: applications, claims, and deaths. NARA delivered each set of entries seperately as a set 20 fixed-width .txt files (3 x 20 = 60 files in total).

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

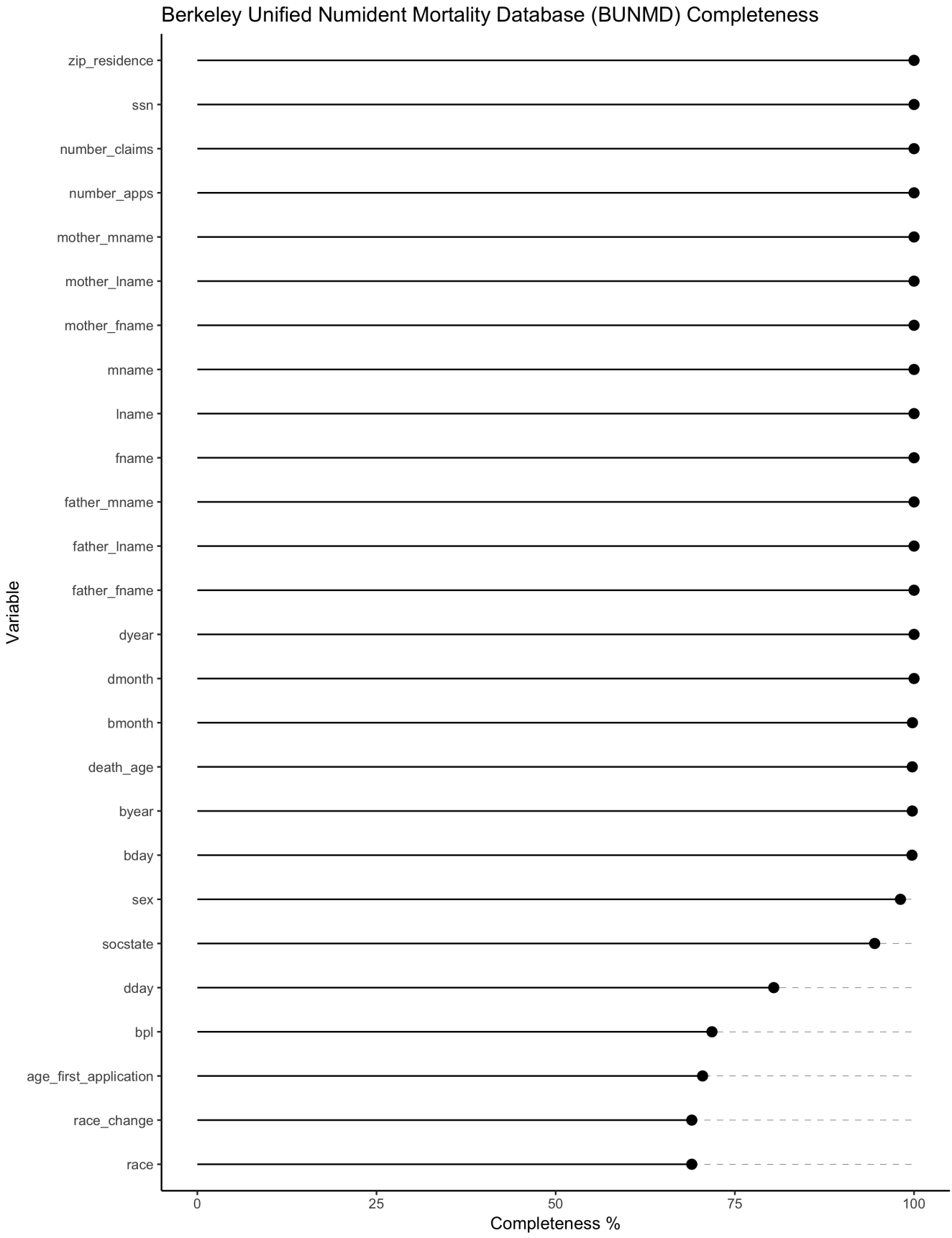
The record layout documents contain variable descriptions, value labels, and other technical notes on the variables in the original NARA Numident file. Additionally, they provide the start and end position for each variables in the fixed-width files.

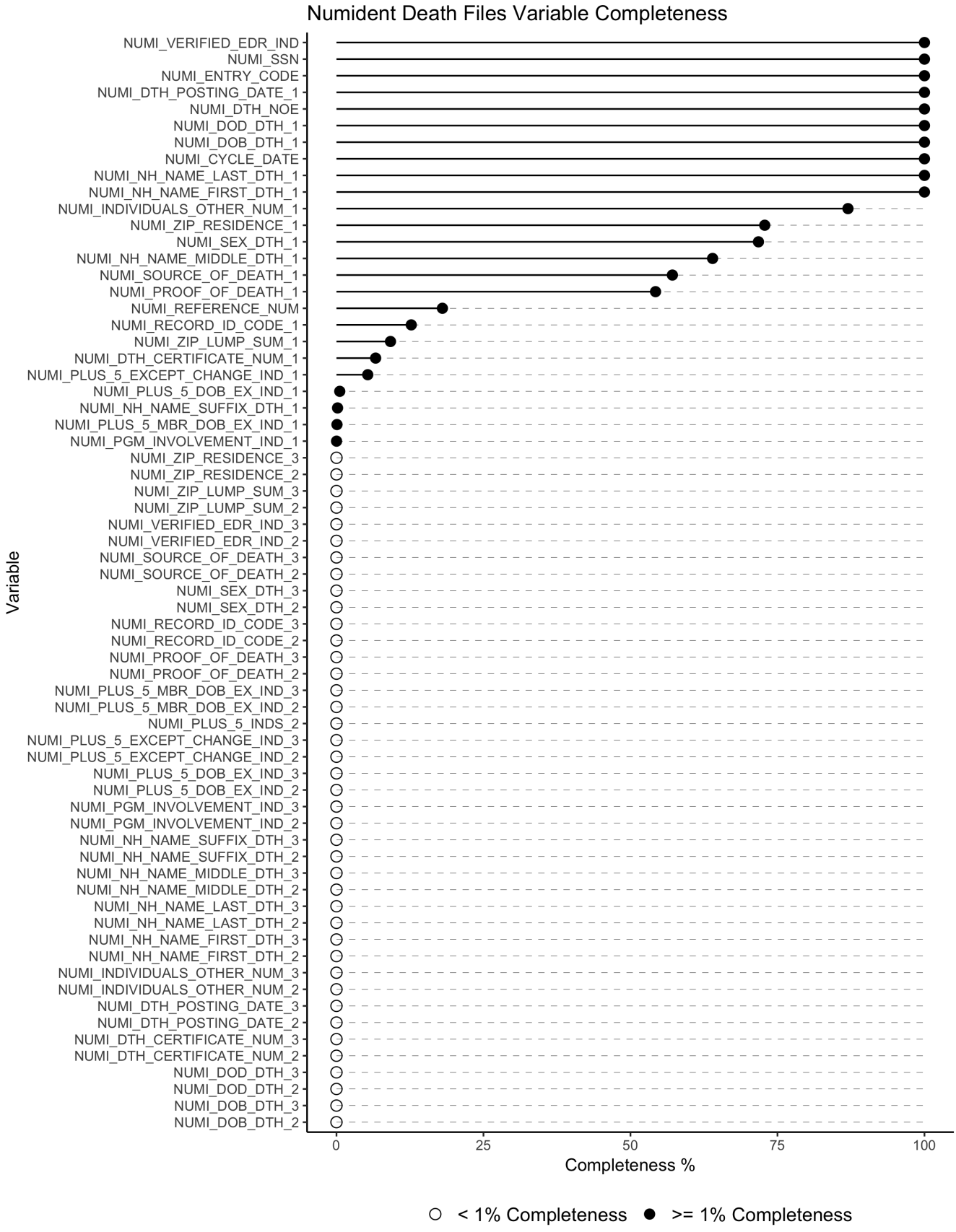
Code

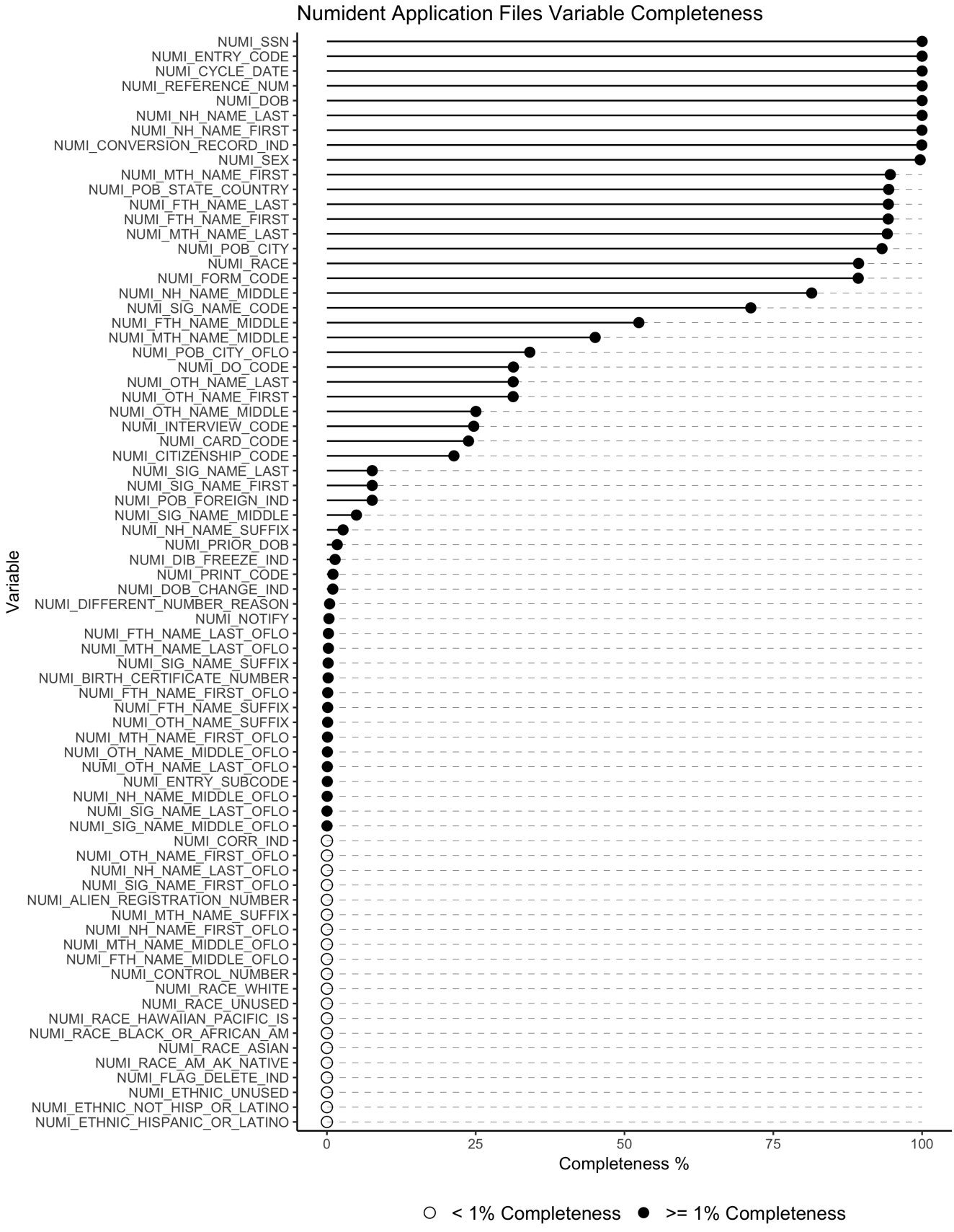
All data processing was was done in the `R` Statistical Programming Language. The code to construct the BUNMD is available at "Github.com/caseybreen/wcensoc/".

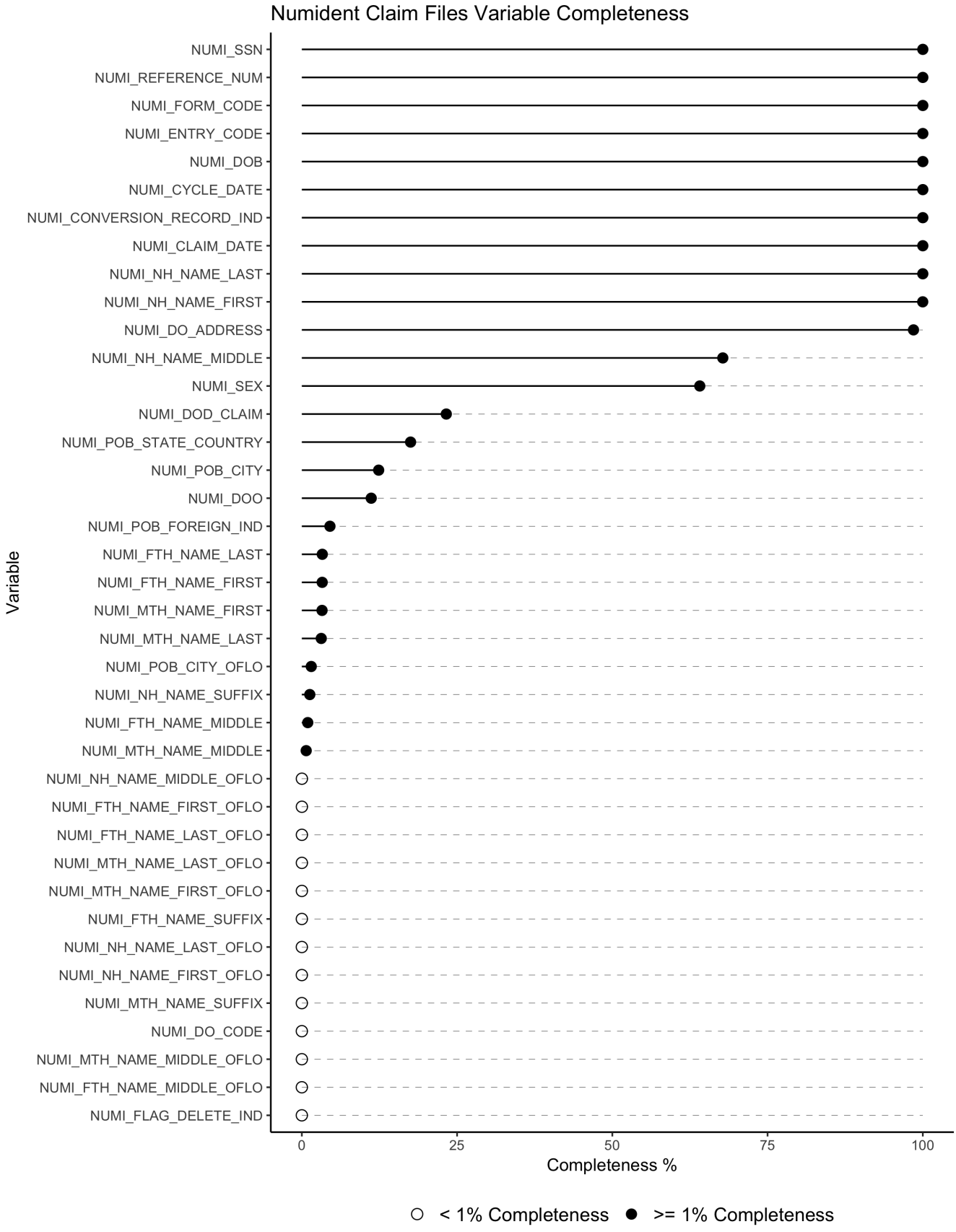
Processing

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


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


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


Numident Application Files Variable Completeness
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
Variable
0
25
50
75
100
Completeness %
< 1% Completeness
>= 1% Completeness


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

file. The end result was three files, one for each of the entry types.

We took the following steps to clean each file:

1. We kept only relevant variables with a low proportion of missing values.
2. We changed the variable names to be more concise and informative. For example, we renamed “NUMI_RACE” variable to “race”.
3. Several values denoted a missing value — “Unkown”, "“," " “Unk”, “Un”, and “0” are all used to represent missing. We replaced these values with “NA.”

Geography

There are several geography 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 of 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. 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.

The Numident death entries contains the 9-digit ZIP code of the residence at the time of death. Sometimes, the full 9-digit ZIP is not available. An “x” represents a missing ZIP code digit. This is the original convention used by the Social Security Administration.

The first three digits of a Social Security number correspond to the state in which a social security card was issued (prior to 1973) or to the ZIP Code of the mailing address listed in the Social Security application (post 1973).

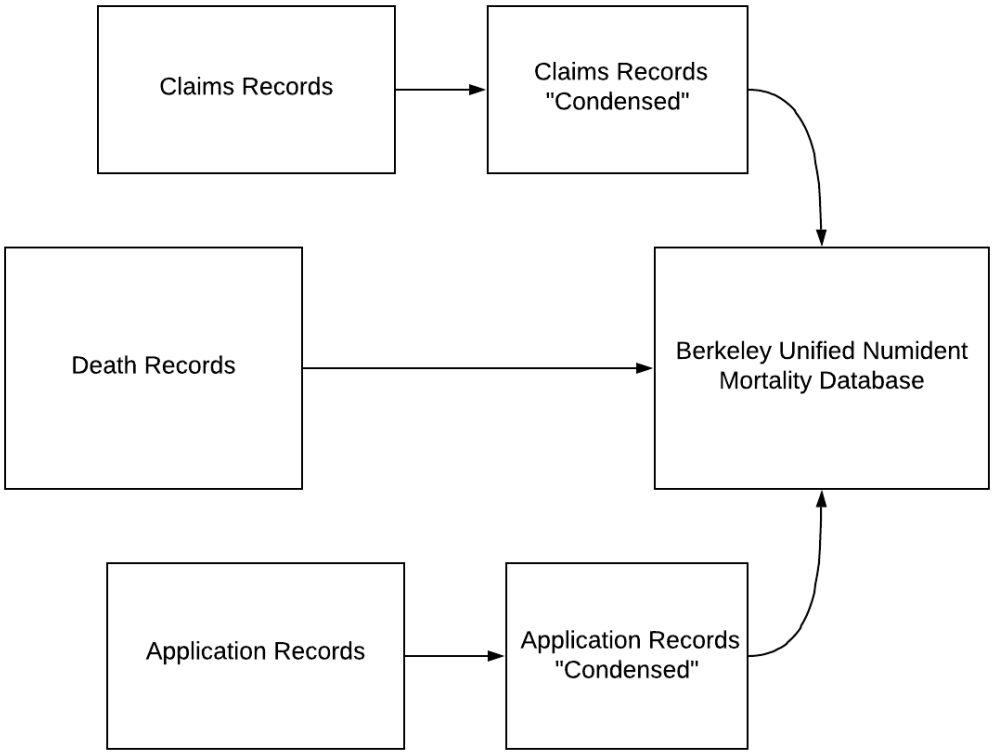


Figure 7: Berkeley Unified Numident Mortality Database creation flowchart.

BUNMD

The BUNMD file contains 28 variables and 49,459,293 records. Using death records not in the high-coverage subsamples of the BUNMD presents challenges. See the BUNMD Codebook for variable descriptions, value labels, and tabulations.

BUNMD Samples and Weights

We create two high death coverage samples from the BUNMD. Sample 1 contains individuals with high coverage: the cohorts of 1900-1940 dying between 1988 and 2005. Sample 2 contains cohorts for 1900-1940 dying between 1988 and 2005 with a valid value for sex, race, and birthplace. We two sets of weights, one for each sample. We constructed a post-stratification weight to the Human Mortality Database (HMD) totals. 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" total.

$$W_j = \frac{\text{HMD deaths in cell j}}{\text{Numident Sample 1 deaths in cell j}} \tag{5}$$

Combining Numident Entries into the BUNMD File

The goal in constructing the BUNMD was to take the information in the NARA Numident files and combine them into a single, harmonized file with one record per person. We made several decisions to reduce the information into a single file. The death files contains a single entry per person. The application and claims records often contain more than one entry per person. We used the a set of decision rules to select the "best" value. Further, values for a given variable may be available in different entries. For example, information on sex is available in the application, claim, and death records. We used the following decision rules to select the "best" value for across all entries:

For persons with multiple application or claim records, there was occasional response inconsistency for sex, race, place of birth, and parents' names. We developed a set of decision rules to select the best value for the Berkeley Unified Numident Mortality Database:

- **sex**: When available, select the person's sex from their death record. If unavailable in their death record, select sex from their most recent application record. If unavailable in their death record and their application records, select sex from their most recent claim record.

- **race**: Select the person's race from their most recent application record.
- **bpl**: Select the person's race from their most recent application record. If unavailable in their application records, select from their most recent claim record.
- **father_fname**: Select the person's father's first name that is the maximum number of characters across applications.
- **father_mname**: Select the person's father's middle name that is the maximum number of characters across applications.
- **father_lname**: Select the person's father's last name that is the maximum number of characters across applications.
- **mother_fname**: Select the person's mother's first name that is the maximum number of characters across applications.
- **mother_mname**: Select the person's mother's middle name that is the maximum number of characters across applications.
- **mother_lname**: Select the person's mother's last name that is the maximum number of characters across applications.

{Original Numident File Source}

Variable	Source
ssn	Death Files
fname	Death Files
mname	Death Files
lname	Death Files
sex	Death, Application, or Claim Files
race	Application Files
race_change	Constructed
bpl	Application or Claim Files
byear	Death Files
bmonth	Death Files
bday	Death Files
dyear	Death Files
dmonth	Death Files
dday	Death Files
death_age	Constructed
zip_residence	Death Files
socstate	Constructed
father_fname	Application or Claim Files
father_mname	Application or Claim Files
father_lname	Application or Claim Files
mother_fname	Application or Claim Files
mother_mname	Application or Claim Files
mother_lname	Application or Claim Files
age_first_app	Constructed
number_apps	Constructed
number_claims	Constructed
weight	Constructed
cweight	Constructed