Supplementary 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 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”.

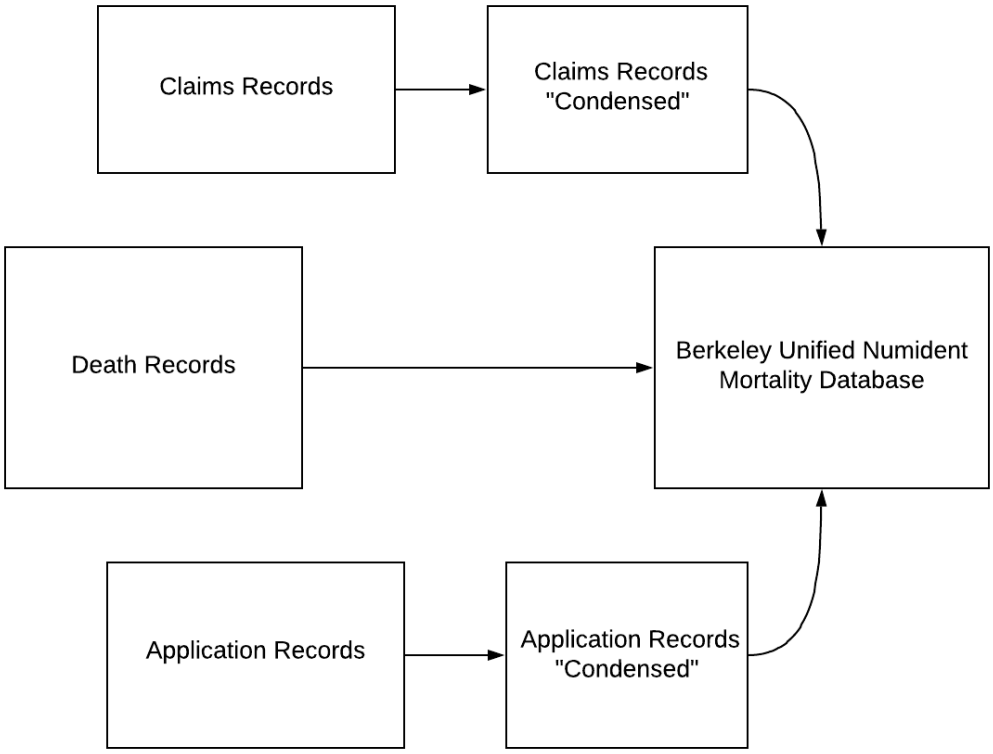
3. Several values denoted a missing value — “Unkown”, "“," " “Unk”, “Un”, and “0” are all used to represent missing. We replaced these values with “NA.”

Geography

There are several geography variables in the NARA Numident records. The Numident application entries has 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).



(a)

Figure 1: Berkeley Unified Numident Mortality Database creation flowchart.

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

BUNMD Samples and Weights

To facilitate mortality research, we created two 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.

$$W_j = \frac{\text{HMD deaths in cell j}}{\text{Numident Sample 1 deaths in cell j}} \qquad \text{for each age.at.death X cohort X sex X year.of.death cell } n \tag{1}$$

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.