Berkeley Unified Numident Mortality Database: Public administrative records for individual-level mortality research

Full Count Census Data II: Record Linkage and Databases

Casey F. Breen ¹  Joshua R. Goldstein ¹

¹University of California, Berkeley | Department of Demography

November 18, 2022
Motivation

- We are far from a complete understanding of the causal determinants of health and mortality in the United States.
Motivation

- We are far from a complete understanding of the causal determinants of health and mortality in the United States.
- Mortality research is often hampered by data limitations.
Motivation

▸ We are far from a complete understanding of the causal determinants of health and mortality in the United States

▸ Mortality research is often hampered by data limitations
  ▸ U.S. has no population-level registry like Scandinavian countries
Motivation

▶ We are far from a complete understanding of the causal determinants of health and mortality in the United States

▶ Mortality research is often hampered by data limitations
  ▶ U.S. has no population-level registry like Scandinavian countries

▶ Researchers are increasingly turning to administrative datasets (Chetty et al., 2016; Card, Dobkin and Maestas, 2008; Card et al., 2010; Meyer and Mittag, 2019; Ruggles, 2014)
The Social Security Numident (Numerical Index) tracks Social Security Number holders.
The Social Security Numident (Numerical Index) tracks Social Security Number holders:

- Date of birth, date of death, birthplace, race, sex, parents names, etc.
The Social Security Numident (Numerical Index) tracks Social Security Number holders

- Date of birth, date of death, birthplace, race, sex, parents' names, etc.

- Internal restricted version used for research by SSA researchers and collaborators (Mehta et al., 2016; Elo et al., 2004; Waldron, 2007)
National Archives public release of Numident records

A subset of Numident records were transferred from the Social Security Administration to the National Archives
National Archives public release of Numident records

- A subset of Numident records were transferred from the Social Security Administration to the National Archives
- National Archives made these records available — 60 text files with 120+ fields with many missing values
- Messy, challenging data structure
## The structure of the National Archives Numident records

<table>
<thead>
<tr>
<th>Record type</th>
<th>Total entries</th>
<th>Total records (persons)</th>
<th>Entries per person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death</td>
<td>49,459,293</td>
<td>49,459,293</td>
<td>1.000</td>
</tr>
<tr>
<td>Applications</td>
<td>72,120,516</td>
<td>40,870,455</td>
<td>1.765</td>
</tr>
<tr>
<td>Claims</td>
<td>25,228,257</td>
<td>25,140,847</td>
<td>1.004</td>
</tr>
</tbody>
</table>
Creating the BUNMD

1. Select names and vital dates from death records
Creating the BUNMD

1. Select names and vital dates from death records

2. Add **harmonized** covariates from application and claim records (and reconcile discrepant values)
Creating the BUNMD

1. Select names and vital dates from death records

2. Add harmonized covariates from application and claim records (and reconcile discrepant values)

3. Create new variables (e.g., age of death, state where SSN was issued)
Creating the BUNMD

1. Select names and vital dates from death records

2. Add **harmonized** covariates from application and claim records (and reconcile discrepant values)

3. Create new variables (e.g., age of death, state where SSN was issued)
### Variables in the BUNMD

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Numident Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssn</td>
<td>Social Security Number</td>
<td>Death Entry</td>
</tr>
<tr>
<td>fname</td>
<td>First name</td>
<td>Death Entry</td>
</tr>
<tr>
<td>mname</td>
<td>Middle name</td>
<td>Death Entry</td>
</tr>
<tr>
<td>lname</td>
<td>Last Name</td>
<td>Death Entry</td>
</tr>
<tr>
<td>byear</td>
<td>Year of birth</td>
<td>Death Entry</td>
</tr>
<tr>
<td>bmonth</td>
<td>Month of birth</td>
<td>Death Entry</td>
</tr>
<tr>
<td>bday</td>
<td>Day of birth</td>
<td>Death Entry</td>
</tr>
<tr>
<td>dyear</td>
<td>Year of death</td>
<td>Death Entry</td>
</tr>
<tr>
<td>dmonth</td>
<td>Month of death</td>
<td>Death Entry</td>
</tr>
<tr>
<td>dday</td>
<td>Day of death</td>
<td>Death Entry</td>
</tr>
<tr>
<td>zip_residence</td>
<td>ZIP Code of residence at death</td>
<td>Death Entry</td>
</tr>
<tr>
<td>sex</td>
<td>Sex</td>
<td>Death, Application, or Claim Entry</td>
</tr>
<tr>
<td>race_first</td>
<td>Race (first)</td>
<td>Application Entry</td>
</tr>
<tr>
<td>race_last</td>
<td>Race (last)</td>
<td>Application Entry</td>
</tr>
<tr>
<td>bpl</td>
<td>Place of birth</td>
<td>Application or Claim Entry</td>
</tr>
<tr>
<td>father_fname</td>
<td>Father’s first name</td>
<td>Application or Claim Entry</td>
</tr>
<tr>
<td>father_mname</td>
<td>Father’s middle name</td>
<td>Application or Claim Entry</td>
</tr>
<tr>
<td>father_lname</td>
<td>Father’s last name</td>
<td>Application or Claim Entry</td>
</tr>
<tr>
<td>mother_fname</td>
<td>Mother’s first name</td>
<td>Application or Claim Entry</td>
</tr>
<tr>
<td>mother_mname</td>
<td>Mother’s middle name</td>
<td>Application or Claim Entry</td>
</tr>
<tr>
<td>mother_lname</td>
<td>Mother’s last name</td>
<td>Application or Claim Entry</td>
</tr>
<tr>
<td>death_age</td>
<td>Age of death (years)</td>
<td>Constructed</td>
</tr>
<tr>
<td>socstate</td>
<td>State in which SS card issued</td>
<td>Constructed</td>
</tr>
<tr>
<td>age_first_app</td>
<td>Age of first application</td>
<td>Constructed</td>
</tr>
<tr>
<td>number_apps</td>
<td>Total number of applications</td>
<td>Constructed</td>
</tr>
<tr>
<td>number_claims</td>
<td>Total number of claims</td>
<td>Constructed</td>
</tr>
<tr>
<td>weight</td>
<td>Weight variable</td>
<td>Constructed</td>
</tr>
<tr>
<td>ccweight</td>
<td>Complete case person-weight</td>
<td>Constructed</td>
</tr>
</tbody>
</table>
Mortality coverage ages 65+


Counts:
- BUNMD
- BUNMD Complete Case
- Human Mortality Database
95%+ mortality coverage between 1988-2005

![Graph showing mortality coverage from 1970 to 2005. The graph compares BUNMD, BUNMD Complete Case, and Human Mortality Database.]
Lexis diagram of death coverage
Double truncation presents challenges for mortality estimation.

**Untruncated**

![Graph showing untruncated data with a difference of 3 years between ages 80 and 83.]

**Doubly truncated**

![Graph showing doubly truncated data with a difference of 1.1 years between ages 84.9 and 86.]

**Key**
- Red: Native-born
- Blue: Foreign-born
Attenuation: Regression understates effects of predictors

\[ \text{Age of Death} = \beta_0 + \lambda_t t + \mathbf{X}\beta + \epsilon \]  \hspace{1cm} (1)

where

1. \( \beta_0 \) is the intercept
Attenuation: Regression understates effects of predictors

Age of Death = $\beta_0 + \lambda_t t + X\beta + \epsilon$  \hspace{1cm} (1)

where

1. $\beta_0$ is the intercept

2. $\lambda_t t$ are birth year fixed effects
Attenuation: Regression understates effects of predictors

Age of Death = $\beta_0 + \lambda_t t + X\beta + \epsilon$ \hspace{1cm} (1)

where

1. $\beta_0$ is the intercept

2. $\lambda_t t$ are birth year fixed effects

3. $X$ is a matrix of covariates and $\beta$ is the coefficient vector
Gompertz parametric MLE approach (no attenuation)

\[ h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \]  

(2)

where

- \( h_i(x|\beta) \) is the hazard at age \( x \) conditional on parameters
**Gompertz parametric MLE approach (no attenuation)**

\[ h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \quad (2) \]

where

- \( h_i(x|\beta) \) is the hazard at age \( x \) conditional on parameters

- \( a_0 \) is some baseline level of mortality
Gompertz parametric MLE approach (no attenuation)

\[ h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \]  

(2)

where

- \( h_i(x|\beta) \) is the hazard at age \( x \) conditional on parameters
- \( a_0 \) is some baseline level of mortality
- \( b_0 \) gives rate of increase of mortality over time
Gompertz parametric MLE approach (no attenuation)

\[ h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \]  

(2)

where

- \( h_i(x|\beta) \) is the hazard at age \( x \) conditional on parameters
- \( a_0 \) is some baseline level of mortality
- \( b_0 \) gives rate of increase of mortality over time
- \( Z_i \) are the covariates for person \( i \)
Gompertz parametric MLE approach (no attenuation)

\[ h_i(x | \beta) = a_0 e^{b_0 x} e^{\beta Z_i} \]  \hspace{1cm} (2)

where

- \( h_i(x | \beta) \) is the hazard at age \( x \) conditional on parameters
- \( a_0 \) is some baseline level of mortality
- \( b_0 \) gives rate of increase of mortality over time
- \( Z_i \) are the covariates for person \( i \) (e.g., years of education, place of birth)
- \( \beta \) is the set of coefficients
Case study 1: Mortality advantage of the foreign born

Introduction

Creating BUNMD

Mortality Estimation

Case Studies

Conclusion

Reserve slides

References
Case study 2: ZIP Code level mortality estimation

Difference in e65

-2.1 to -1.3
-1.3 to -0.1
-0.1 to 0.1
0.1 to 0.3
0.3 to 0.6
0.6 to 1.2
1.2 to 13.6
NA

Avg. SS Benefits

809 to 909
909 to 956
956 to 1,022
1,022 to 1,036
1,036 to 1,077
1,077 to 1,114
1,114 to 1,308

Introduction
Creating BUNMD
Mortality Estimation
Case Studies
Conclusion
Reserve slides
References
CenSoc-Numident: Linking BUNMD with the 1940 Census

IPUMS USA Full-Count 1940 Census

Berkeley Unified Numident Mortality Database (Deaths from 1988 - 2005)

ABE Exact Match on: first name, last name, place of birth, and census age

CenSoc-Numident (N = 8 million)
Considerations and future directions

- **Problem** Double truncation can downwardly attenuate estimates from conventional regression
Considerations and future directions

- **Problem** Double truncation can downwardly attenuate estimates from conventional regression

- **Solution**: Parametric maximum likelihood methods (Gompertztrunc R Package)

Identify siblings using parents names and machine learning techniques (Joo et al.)

Link onto 1950 Census, WWII enlistment records
Considerations and future directions

- **Problem**: Double truncation can downwardly attenuate estimates from conventional regression
  - **Solution**: Parametric maximum likelihood methods (Gompertztrunc R Package)

- Identify siblings using parents names and machine learning techniques (Joo et al.)
Considerations and future directions

▶ **Problem** Double truncation can downwardly attenuate estimates from conventional regression

▶ **Solution:** Parametric maximum likelihood methods *(Gompertztrunc R Package)*

▶ Identify siblings using parents names and machine learning techniques *(Joo et al.)*

▶ Link onto 1950 Census, WWII enlistment records
Conclusions

- BUNMD: publicly available file containing 50 million mortality records and covariates
Conclusions

- BUNMD: publicly available file containing 50 million mortality records and covariates
- Linked onto the 1940 Census ($N = 9$ million)
Conclusions

- BUNMD: publicly available file containing 50 million mortality records and covariates

- Linked onto the 1940 Census ($N = 9$ million)

- **Publicly Available:** Reproducible, extendable science. No barriers to entry.
Thank You

Download: CenSoc.Berkeley.edu

Funding: R01AG058940, R01AG076830

Contact: caseybreen@berkeley.edu

DEMOPGRAPHIC RESEARCH

VOLUME 47, ARTICLE 5, PAGES 111–142
PUBLISHED 14 JULY 2022
http://www.demographic-research.org/Volumes/Vol47/5/
DOI: 10.4054/DemRes.2022.47.5

Research Material

Berkeley Unified Numident Mortality Database: Public administrative records for individual-level mortality research

Casey F. Breen

Joshua R. Goldstein
González et al. — Hispanic mortality paradox

![Graph showing relative life expectancy for women across different countries compared to Native White Movers.](image-url)
## Goldstein et al. — Black names and longevity

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>Pooled</th>
<th>Death Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(3)</td>
</tr>
<tr>
<td><strong>Model:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BNI (Standardized)</td>
<td>-0.2386</td>
<td>-0.6258*</td>
</tr>
<tr>
<td></td>
<td>(0.2301)</td>
<td>(0.3060)</td>
</tr>
<tr>
<td>Birth Year FE</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Family FE</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>Birth Order FE</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Observations</strong></td>
<td>30,429</td>
<td>30,429</td>
</tr>
<tr>
<td><strong>R^2</strong></td>
<td>0.21029</td>
<td>0.61428</td>
</tr>
<tr>
<td><strong>Within R^2</strong></td>
<td>5.35 × 10^{-5}</td>
<td>0.00036</td>
</tr>
</tbody>
</table>

**Note:** *p<0.1; **p<0.05; ***p<0.01
### Creating BUNMD

#### Mortality Estimation

<table>
<thead>
<tr>
<th></th>
<th>ML, TH1</th>
<th>ML, TH2</th>
<th>EM</th>
<th>EM, within birthplace</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Weighted mean</td>
<td>Mean</td>
<td>Weighted mean</td>
</tr>
<tr>
<td>Total</td>
<td>1.53</td>
<td>1.35</td>
<td>1.31</td>
<td>1.14</td>
</tr>
<tr>
<td>Race: White</td>
<td>1.43</td>
<td>1.26</td>
<td>1.23</td>
<td>1.07</td>
</tr>
<tr>
<td>Race: Black</td>
<td>2.33</td>
<td>2.01</td>
<td>1.85</td>
<td>1.57</td>
</tr>
<tr>
<td>Race: Others</td>
<td>2.15</td>
<td>1.92</td>
<td>1.78</td>
<td>1.55</td>
</tr>
<tr>
<td>Cohort: 1900-4</td>
<td>0.77</td>
<td>0.74</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>Cohort: 1905-9</td>
<td>1.15</td>
<td>1.11</td>
<td>0.92</td>
<td>0.88</td>
</tr>
<tr>
<td>Cohort: 1910-4</td>
<td>1.44</td>
<td>1.42</td>
<td>1.21</td>
<td>1.19</td>
</tr>
<tr>
<td>Cohort: 1915-9</td>
<td>1.56</td>
<td>1.56</td>
<td>1.35</td>
<td>1.34</td>
</tr>
<tr>
<td>Cohort: 1920-4</td>
<td>1.65</td>
<td>1.64</td>
<td>1.42</td>
<td>1.42</td>
</tr>
<tr>
<td>Cohort: 1925-9</td>
<td>1.61</td>
<td>1.61</td>
<td>1.38</td>
<td>1.38</td>
</tr>
<tr>
<td>Cohort: 1930-4</td>
<td>1.47</td>
<td>1.46</td>
<td>1.24</td>
<td>1.23</td>
</tr>
<tr>
<td>% of any sibling</td>
<td>53.8%</td>
<td>50.6%</td>
<td>27.7%</td>
<td>25.0%</td>
</tr>
</tbody>
</table>
## Variable source and selection rule

<table>
<thead>
<tr>
<th>Variable</th>
<th>Numident source</th>
<th>Selection rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssn</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>fname</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>mname</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>lname</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>byear</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>bmonth</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>bday</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>dyear</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>dmonth</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>dday</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>zip_residence</td>
<td>Death Entry</td>
<td>-</td>
</tr>
<tr>
<td>sex</td>
<td>Death, Application, or Claim Entry</td>
<td>Last Recorded Sex</td>
</tr>
<tr>
<td>race_first</td>
<td>Application Entry</td>
<td>First Recorded Race</td>
</tr>
<tr>
<td>race_last</td>
<td>Application Entry</td>
<td>Last Recorded Race</td>
</tr>
<tr>
<td>bpl</td>
<td>Application or Claim Entry</td>
<td>Last Recorded BPL</td>
</tr>
<tr>
<td>father_fname</td>
<td>Application or Claim Entry</td>
<td>Maximum Characters</td>
</tr>
<tr>
<td>father_mname</td>
<td>Application or Claim Entry</td>
<td>Maximum Characters</td>
</tr>
<tr>
<td>father_lname</td>
<td>Application or Claim Entry</td>
<td>Maximum Characters</td>
</tr>
<tr>
<td>mother_fname</td>
<td>Application or Claim Entry</td>
<td>Maximum Characters</td>
</tr>
<tr>
<td>mother_mname</td>
<td>Application or Claim Entry</td>
<td>Maximum Characters</td>
</tr>
<tr>
<td>mother_lname</td>
<td>Application or Claim Entry</td>
<td>Maximum Characters</td>
</tr>
<tr>
<td>death_age</td>
<td>Constructed</td>
<td>-</td>
</tr>
<tr>
<td>socstate</td>
<td>Constructed</td>
<td>-</td>
</tr>
<tr>
<td>age_first_app</td>
<td>Constructed</td>
<td>-</td>
</tr>
<tr>
<td>number_apps</td>
<td>Constructed</td>
<td>-</td>
</tr>
<tr>
<td>number_claims</td>
<td>Constructed</td>
<td>-</td>
</tr>
<tr>
<td>weight</td>
<td>Constructed</td>
<td>-</td>
</tr>
<tr>
<td>ccweight</td>
<td>Constructed</td>
<td>-</td>
</tr>
</tbody>
</table>
Linking unmarried women in CenSoc-Numident

Sarah Johnston changes her name to Sarah Smith after she is observed in 1940 census

Census Record (1940)  
Sarah Johnston

Marriage (1944)  
Name Change

BUNMD Record (1995)  
Sarah Smith

We can still establish a match using father's last name from Numident Record.

1940 Census Record  
Name for linkage: Sarah Johnston (unmarried)

Match Established

Name: Sarah Smith  
Father's name: Jim Johnston  
Name for linkage: Sarah Johnston

BUNMD Mortality Record

No Match Established

Name: Sarah Smith  
Father's name: Jim Johnston  
Name for linkage: Sarah Smith

References
References


