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]

[1]University 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)

Numident: Backbone of SSA record keeping system

- The Social Security Numident (Numerical Index) tracks Social Security Number holders

Numident: Backbone of SSA record keeping system

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

Numident: Backbone of SSA record keeping system

- 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

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

Creating the BUNMD

1. Select names and vital dates from death records

Creating the BUNMD

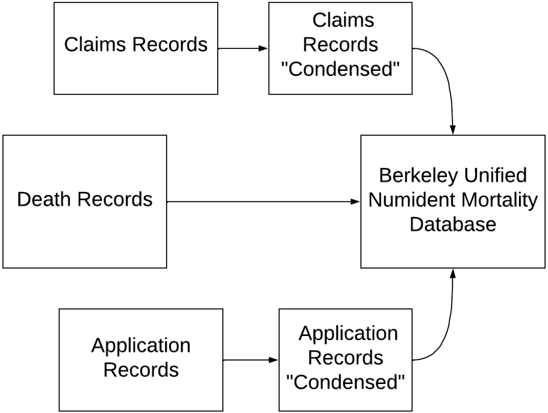
1. Select names and vital dates from death records
2. Add **harmonized** covariates from application and claim records (and reconcile discrepant values)

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

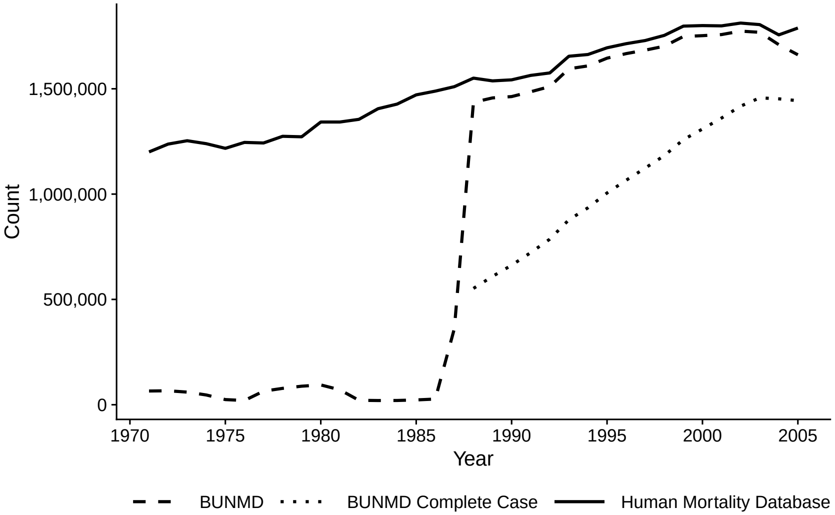
1. Select names and vital dates from death records
2. Add **harmonized** covariates from application and claim records (and reconcile discrepant values)
3. Create new variables (e.g., age of death, state where SSN was issued)



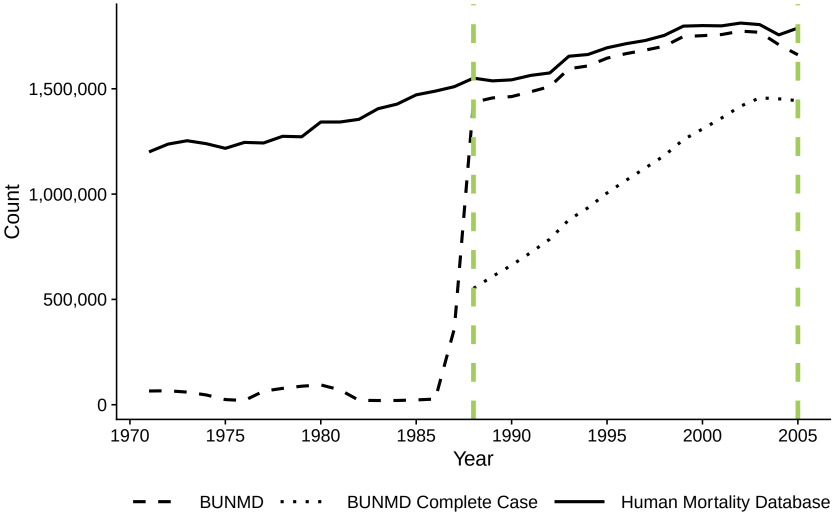
Variables in the BUNMD

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
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

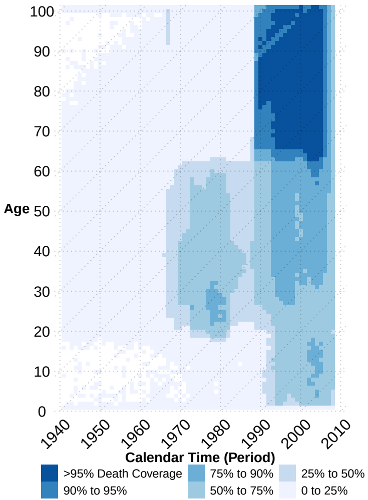
Mortality coverage ages 65+



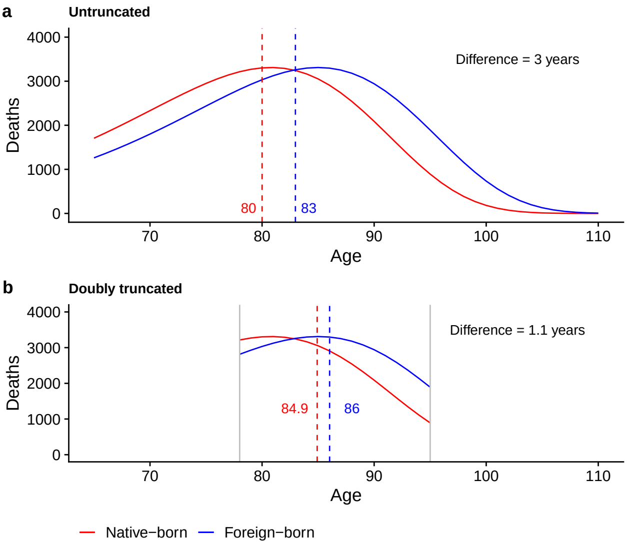
95%+ mortality coverage between 1988-2005



Lexis diagram of death coverage



Double truncation presents challenges for mortality estimation



Attenuation: Regression understates effects of predictors

$$\text{Age of Death} = \beta_0 + \lambda_t t + \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{1}$$

where

1. β_0 is the intercept

Attenuation: Regression understates effects of predictors

$$\text{Age of Death} = \beta_0 + \lambda_t t + \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{1}$$

where

1. β_0 is the intercept
2. $\lambda_t t$ **are birth year fixed effects**

Attenuation: Regression understates effects of predictors

$$\text{Age of Death} = \beta_0 + \lambda_t t + \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{1}$$

where

1. β_0 is the intercept
2. $\lambda_t t$ **are birth year fixed effects**
3. $\mathbf{X}$ **is a matrix of covariates and** $\boldsymbol{\beta}$ **is the coefficient vector**

Gompertz parametric MLE approach (no attenuation)

$$h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \tag{2}$$

where

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

Gompertz parametric MLE approach (no attenuation)

$$h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \tag{2}$$

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)

$$h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \tag{2}$$

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)

$$h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \tag{2}$$

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

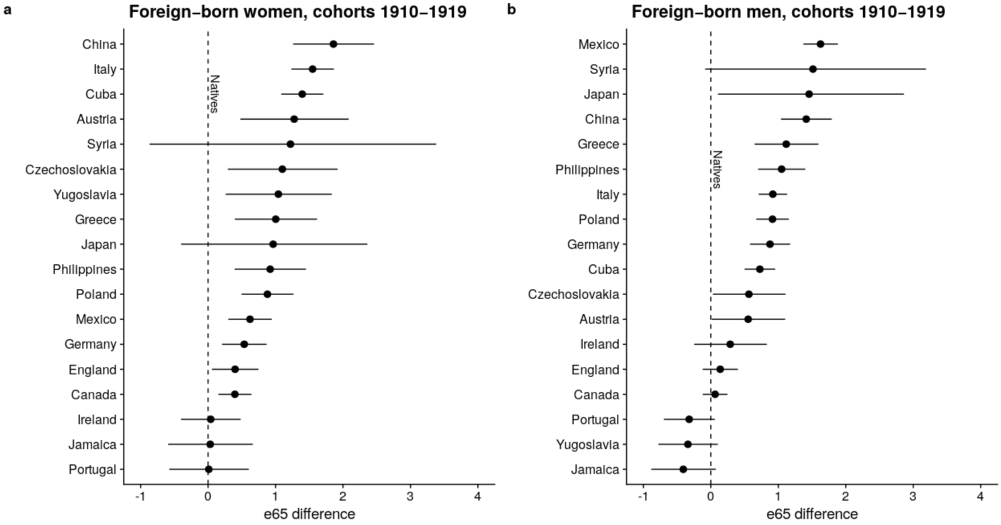
Gompertz parametric MLE approach (no attenuation)

$$h_i(x|\beta) = a_0 e^{b_0 x} e^{\beta Z_i} \tag{2}$$

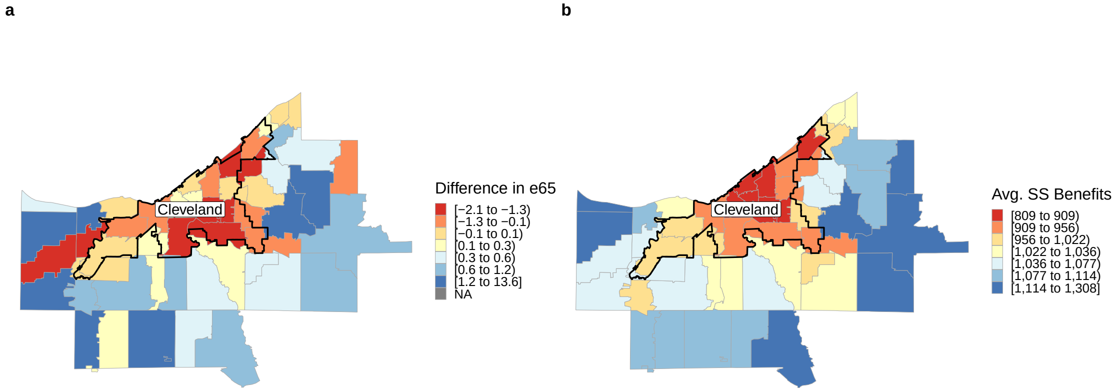
where

- $h_i(x|\beta)$ 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)
- β is the set of coefficients

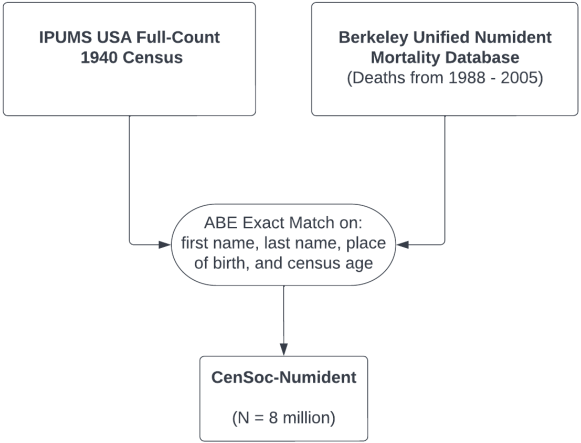
Case study 1: Mortality advantage of the foreign born



Case study 2: ZIP Code level mortality estimation



CenSoc-Numident: Linking BUNMD with the 1940 Census



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)

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

DEMOGRAPHIC 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

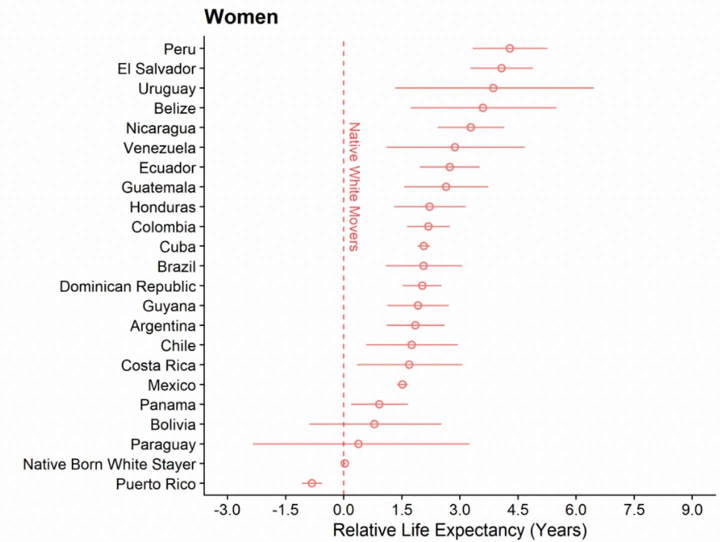
Berkeley Unified Numident Mortality Database: Public administrative records for individual-level mortality research

Casey F. Breen

Joshua R. Goldstein

Reserve Slides

González et al. — Hispanic mortality paradox



Goldstein et al. — Black names and longevity

Dependent Variable:	Death Age			
	Pooled	**Family FE**		
Model:	(1)	(3)	(4)	(5)
BNI (Standardized)	-0.2386	-0.6258*	-0.6273*	-0.4696
	(0.2301)	(0.3060)	(0.3055)	(0.4380)
Birth Year FE	Yes	Yes	Yes	Yes
Family FE	-	Yes	Yes	Yes
Birth Order FE	-	-	Yes	Yes
Mortality Window	1988-2005	1988-2005	1988-2005	1941-2007
Observations	30,429	30,429	30,429	45,893
R^2	0.21029	0.61428	0.61430	0.56402
Within R^2	5.35×10^{-5}	0.00036	0.00036	8.14×10^{-5}

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

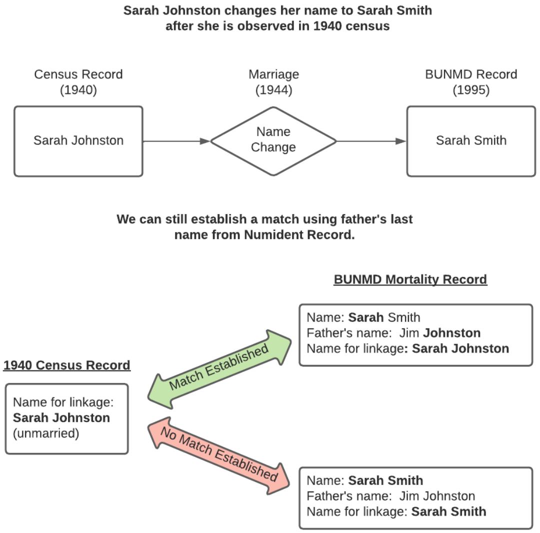
Joo et al. — Identifying siblings using machine learning

	ML, TH1		ML, TH2		EM		EM, within birthplace	
	Mean	Weighted mean	Mean	Weighted mean	Mean	Weighted mean	Mean	Weighted mean
Total	1.53	1.35	1.31	1.14	0.53	0.46	0.44	0.39
Race: White	1.43	1.26	1.23	1.07	0.54	0.48	0.46	0.40
Race: Black	2.33	2.01	1.85	1.57	0.47	0.40	0.38	0.31
Race: Others	2.15	1.92	1.78	1.55	0.23	0.21	0.19	0.17
Cohort: 1900-4	0.77	0.74	0.60	0.57	0.23	0.19	0.21	0.17
Cohort: 1905-9	1.15	1.11	0.92	0.88	0.33	0.28	0.31	0.26
Cohort: 1910-4	1.44	1.42	1.21	1.19	0.47	0.41	0.45	0.39
Cohort: 1915-9	1.56	1.56	1.35	1.34	0.55	0.49	0.55	0.48
Cohort: 1920-4	1.65	1.64	1.42	1.42	0.58	0.51	0.58	0.52
Cohort: 1925-9	1.61	1.61	1.38	1.38	0.55	0.49	0.55	0.49
Cohort: 1930-4	1.47	1.46	1.24	1.23	0.49	0.44	0.49	0.44
% of any sibling	53.8%		50.6%		27.7%		25.0%	

Variable source and selection rule

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
death_age	Constructed	-
socstate	Constructed	-
age_first_app	Constructed	-
number_apps	Constructed	-
number_claims	Constructed	-
weight	Constructed	-
ccweight	Constructed	-

Linking unmarried women in CenSoc-Numident



References

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." *American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas.* .

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.

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.

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.

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

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.