

CenSoc: Public Linked Administrative Mortality Records for Individual-Level Research

New Linked Data Sources

Casey F. Breen^{1,2} Maria Osborne¹ Joshua R. Goldstein¹

¹University of California, Berkeley | Department of Demography

²University of Oxford | Leverhulme Centre for Demographic Science

July 5, 2024



Motivation for CenSoc

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

Motivation for CenSoc

- ▶ 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**

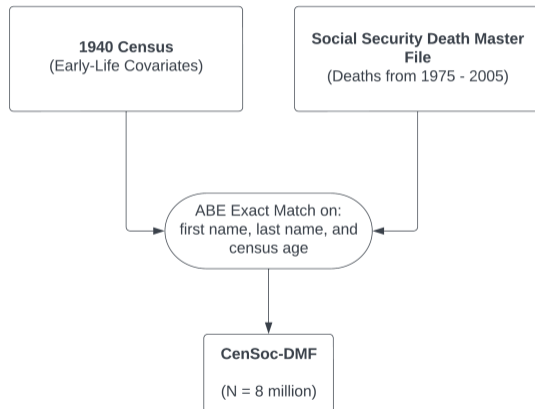
Motivation for CenSoc

- ▶ 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

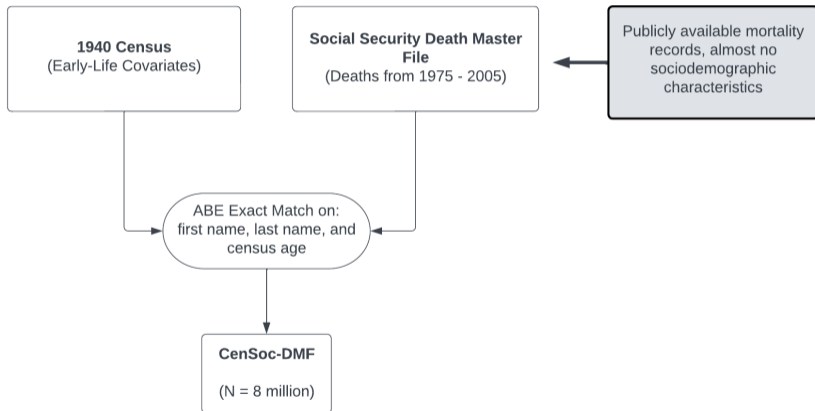
Motivation for CenSoc

- ▶ 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
- ▶ Social scientists are increasingly turning to administrative datasets ([Ruggles, 2014](#); [Chetty et al., 2016](#); [Card et al., 2010](#))

CenSoc: Linked IPUMS 1940 Census and Mortality Records



CenSoc: Linked IPUMS 1940 Census and mortality records



Social Security Mortality Records – Numident



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

Social Security Mortality Records – Numident



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

Social Security Mortality Records – Numident



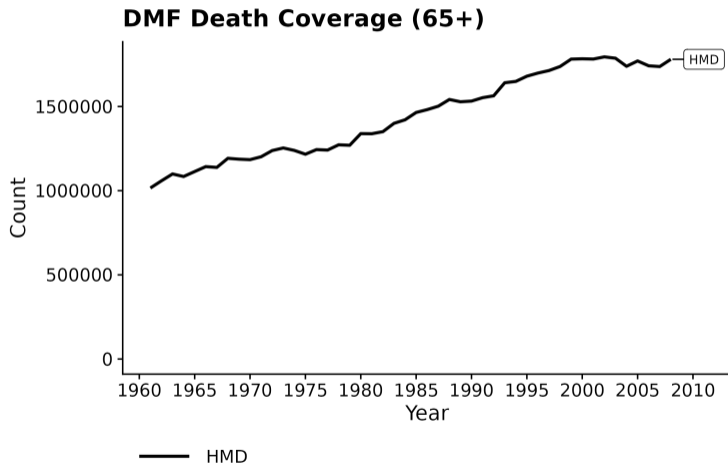
- ▶ The Social Security Numident (Numerical Index) tracks Social Security Number holders
 - ▶ Date of birth, date of death, birthplace, race, sex, parents names, etc.

- ▶ Internal restricted version used for research by SSA researchers and collaborators ([Mehta et al., 2016](#); [Elo et al., 2004](#); [Waldron, 2007](#))

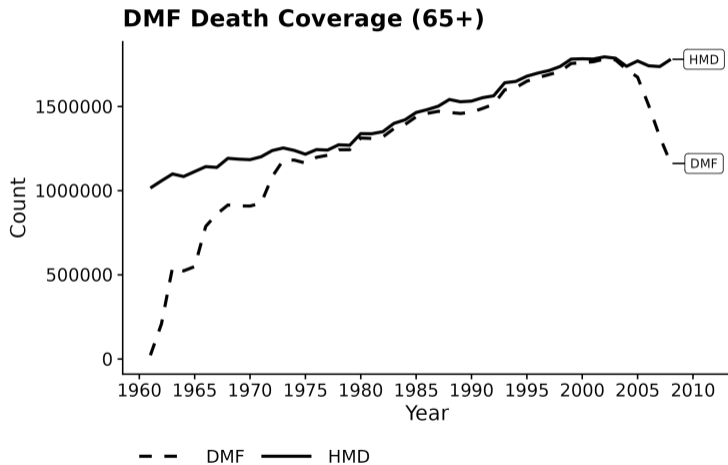
Social Security Mortality Records – Death Master File

- ▶ Social Security Death Master File Death Master File (DMF) is an extract of Numident, plus misc. deaths
- ▶ **Limited info:** Name, date of birth, date of death

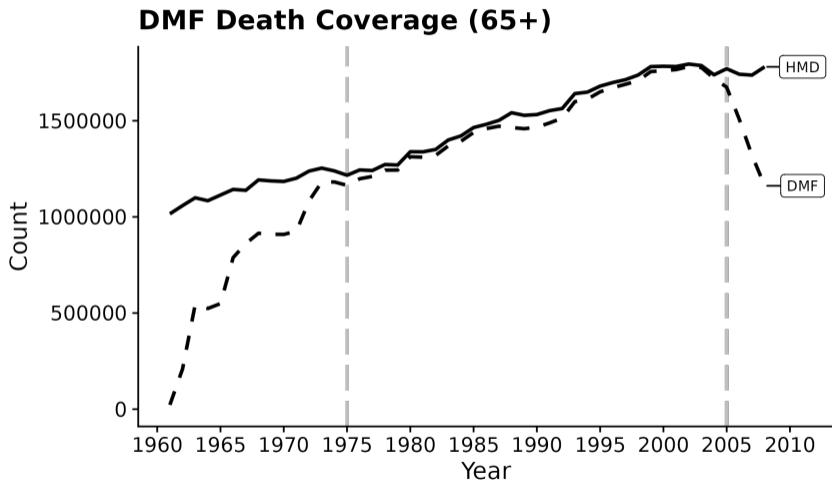
Coverage DMF (Public)



Coverage DMF (Public)

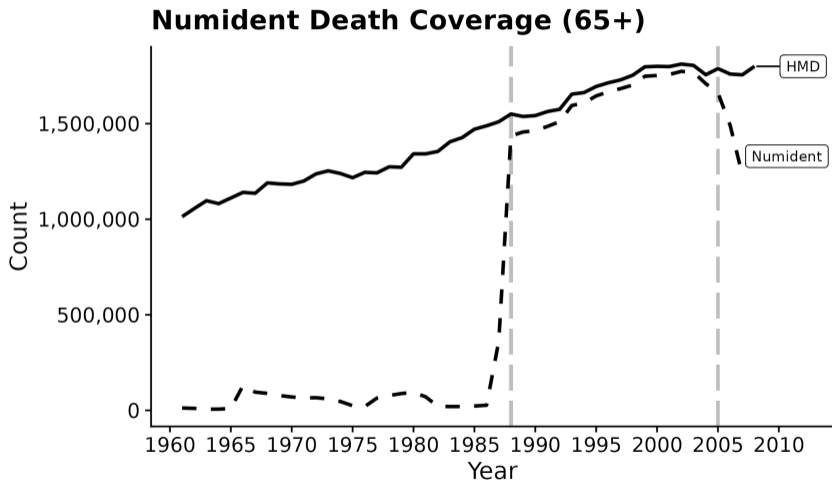


Public DMF — 95% death coverage 1975-2005

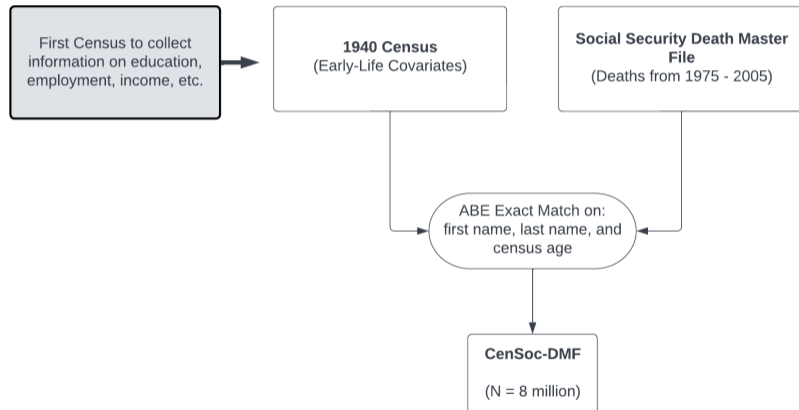


-- DMF — HMD

Public Numident: 95%+ mortality coverage between 1988-2005



Linked IPUMS 1940 Census and mortality records



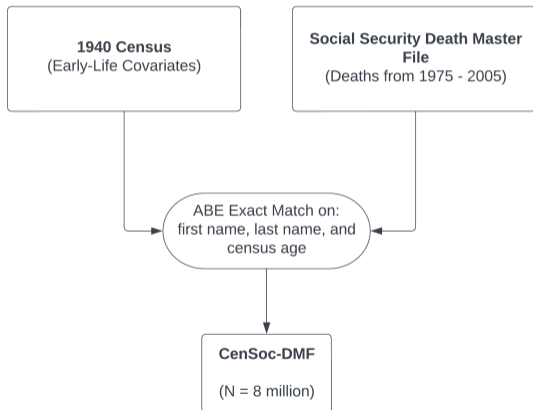
1940 Census

- ▶ 1940 Census reflected heightened time of social awareness brought about by Great Depression
- ▶ First decennial census to include question on educational attainment, wage and salary income, and detailed questions on employment
- ▶ Question on homeownership status (rent vs. own) and estimate of home value for owners

The image shows a detailed 1940 Census form with multiple columns and rows of data. A red oval highlights a specific row, likely representing an individual's record. The form includes various fields for personal information, household details, and economic data.

1940 Census Form

ABE Conservative Algorithm for Record Linkage



Summary of datasets

	CenSoc-DMF	CenSoc-Numident
Gender	Men-Only	Men and Women
1940 Census Covariates	Yes	Yes
Death Coverage	1975-2005	1988-2005
Size	4.7 Million	7.0 Million

Characteristics of CenSoc Datasets

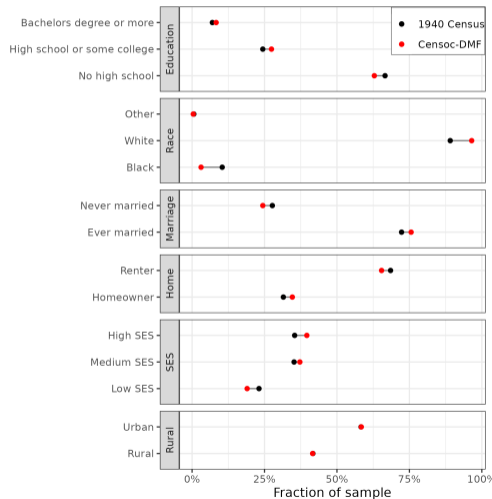
Match rate (mortality adjusted)

$$M_{adjusted} = \underbrace{\left(\frac{\text{Number Established Matches}}{\text{Number of Records in 1940 Census}} \right)}_{\text{Raw match rate}} \times \underbrace{\left(\frac{1}{P(\text{Dying in window})} \right)}_{\text{Adjustment factor for mortality}},$$

- ▶ CenSoc-Numident: 22%
- ▶ CenSoc-DMF: 17%

Mostly representative of general population

- ▶ Compared to the general population, CenSoc is:
 - ▶ Slightly higher socioeconomic status
 - ▶ Slightly more white



Statistical Person-level Weights

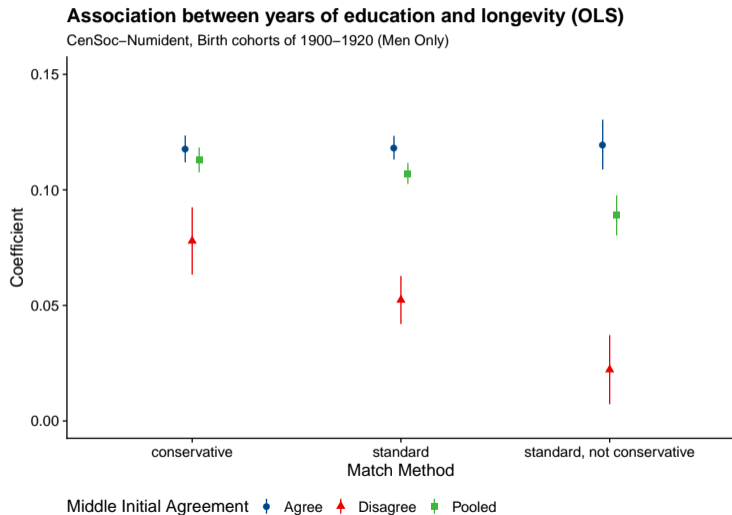
- ▶ **Post-stratification weights:** Use population totals from the Multiple Cause-of-Death (MCOB) mortality data from National Center for Health Statistics (NCHS)
- ▶ Individuals are split into cells cross-classified by year of death (y), age at death (a), sex (s), race (r), and birth state (b)

$$W_{yasrb} = \frac{\text{number of deaths in NCHS cell } yasrb}{\text{number of deaths in CenSoc cell } yasrb}$$

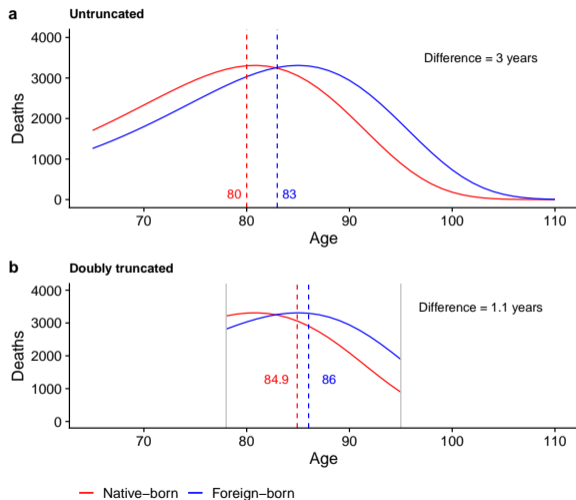
Validation Variable - Middle Initial

- ▶ Middle initials not used as a linking field
- ▶ Can give us upper-bound on false match rate (type 1 error)
- ▶ **Agreement**
 - ▶ 87% CenSoc-Numident
 - ▶ 85% CenSoc-DMF

Validation Variable - Middle Initial



Double truncation presents challenges for mortality estimation



Method 1: OLS regression on age of death (attenuated)

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

where

1. β_0 is the intercept

Method 1: OLS regression on age of death (attenuated)

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

where

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

Method 1: OLS regression on age of death (attenuated)

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

where

1. β_0 is the intercept
2. $\lambda_t t$ are birth year fixed effects
3. X is a matrix of covariates and β is the coefficient vector

Method 2: Gompertz parametric MLE approach (no attenuation)

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

where

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

Method 2: Gompertz parametric MLE approach (no attenuation)

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

where

- ▶ $h_i(x|\beta)$ is the hazard at age x conditional on parameters
- ▶ a_0 is some baseline level of mortality

Method 2: Gompertz parametric MLE approach (no attenuation)

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

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

Method 2: Gompertz parametric MLE approach (no attenuation)

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

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

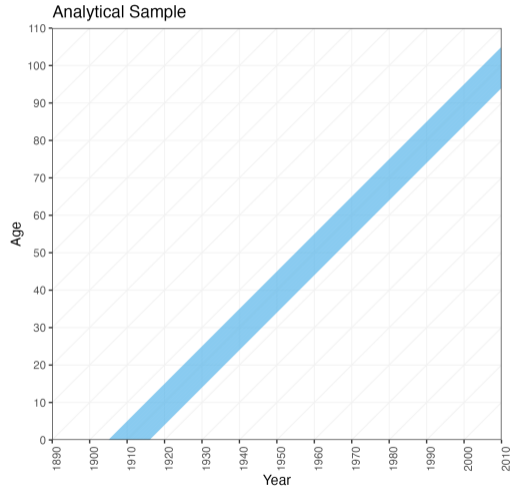
Method 2: Gompertz parametric MLE approach (no attenuation)

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

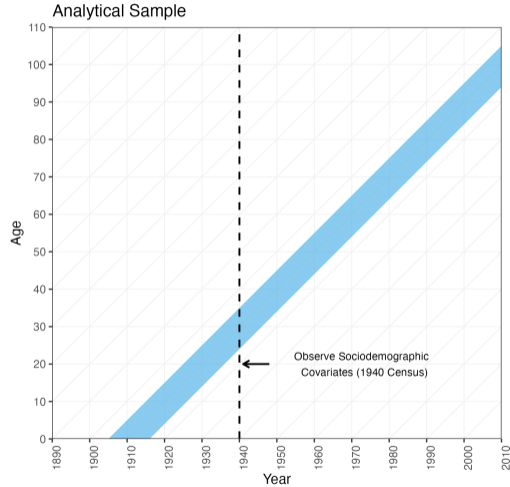
where

- ▶ $h_i(x|\beta)$ 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

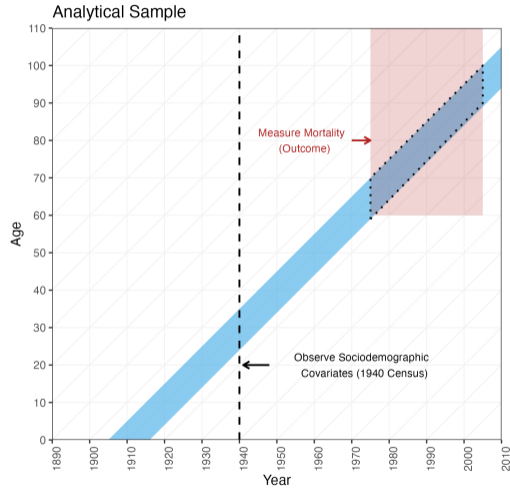
Cohort perspective



Cohort perspective



Cohort perspective



Utility for mortality research

- ▶ Mortality disparities by education, national origin, and race
- ▶ Early life conditions and later-life mortality
- ▶ Geographic variation and the neighborhood determinants of mortality
- ▶ Natural experiments from local policies and chance events such as natural disasters.

Publications

Berkeley Unified Numident Mortality Database: Public Administrative Records for Individual Level Mortality Research. Demographic Research, 47-5, 111-142
Joshua R. Goldstein, Casey F. Green, February 2022

Abstract: With the release of Social Security application (SS-A), claim, and death records, the National Archives and Records Administration (NARS) has created a new administrative 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 population-level register data is a...

Later-life Changes in Ethnoracial Self-Identification: Evidence from Social Security Administrative Data. Population Research and Policy Review, 42: 10
Casey F. Green, September 2022

Abstract: Researchers generally recognize that ethnoracial identification may shift over the life course. However, the prevalence of these shifts across cohorts and among older adults remain open questions. Using administrative data from Social Security applications from 1984 to 2007, we quantify the magnitude and direction of later-life shifts in ethnoracial self-identification among Black, White, ...

Social Insurance Programs and Later-Life Mortality: Evidence from New Deal Relief Spending. Journal of Health Economics, 86
Harold Noghibehbahar, Michal Engelman, December 2022

Abstract: A growing body of research explores the long-run effects of social programs and welfare spending. However, evidence linking welfare support in early life with longevity is limited. We add to this literature by evaluating the effect of in-utero and early-life exposure to the largest increases in welfare spending in the US history under the New Deal ...

Does a Prolonged Hardship Reduce Life Span? Examining the Longevity of Young Men who Lived through the 1930s Great Plains Drought. Population and Environment, 43, 530-552
Serge Atherwood, May 2022

Abstract: The Great Plains drought of 1931-1939 was a prolonged socio-ecological disaster with widespread impacts on society, economy, and health. While its immediate impacts are well-documented, we know much less about the disaster's effects on vital human outcomes. In particular, the event's effects on later-life mortality remain almost entirely unexplored. Closing this gap ...

In utero exposure to natural disasters and later-life mortality: Evidence from earthquakes in the early twentieth century. Social Science & Medicine, 307
Harold Noghibehbahar, August 2022

A growing body of research explores the effects of prenatal insults caused by natural disasters on life-cycle outcomes. This paper joins the literature by exploring the long-run effects of prenatal exposure to earthquakes on adulthood and old-age mortality. Using Social Security Administration death records (1975-2005) linked with the full-count 1940 US census and implementing a ...

Mortality Modeling of Partially Observed Cohorts Using Administrative Death Records
Joshua R. Goldstein, Maria Osborne, Serge Atherwood & Casey F. Green, 26 April 2023

Abstract: New advances in data linkage provide mortality researchers with access to administrative datasets with millions of mortality records and rich demographic covariates. Although these new datasets allow for high-resolution mortality research, administrative mortality records often have technical limitations, such as limited mortality coverage windows and incomplete observation of survivors. We describe a method for ...

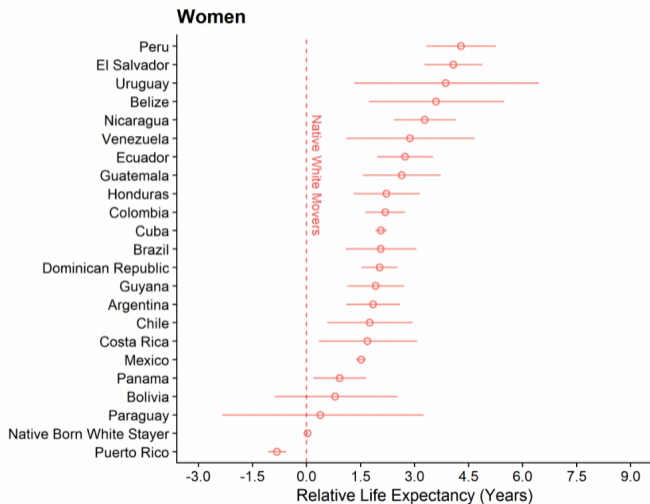
The Early Bird Catches the Worm: The Effect of Birth Order on Old-Age Mortality
Harold Noghibehbahar & Jason Fletcher, July 2023

Abstract: Previous studies explore the role of birth order in children's and adults' outcomes. This literature usually provides evidence of disadvantage of children with higher birth order. In a new strand of this literature, we explore the association between birth order and old-age mortality. This study re-examines the birth-order longevity relationship using US data. We employ Social Security ...

Early life exposure to cigarette smoking and adult and old-age male mortality: Evidence from linked US full-count census and mortality data
James Hefferts & John Robert Warren, October 2023

Abstract: BACKGROUND: Smoking is a leading cause of premature death across contemporary developed nations, but longitudinal individual-level studies have examined the long-term health consequences of exposure to smoking. OBJECTIVE: We examine the effect of total and in-utero exposure to exogenous variation in smoking, brought about by state-level cigarette taxation, on adulthood and old-age mortality (aged 70) among cohorts of boys born ...

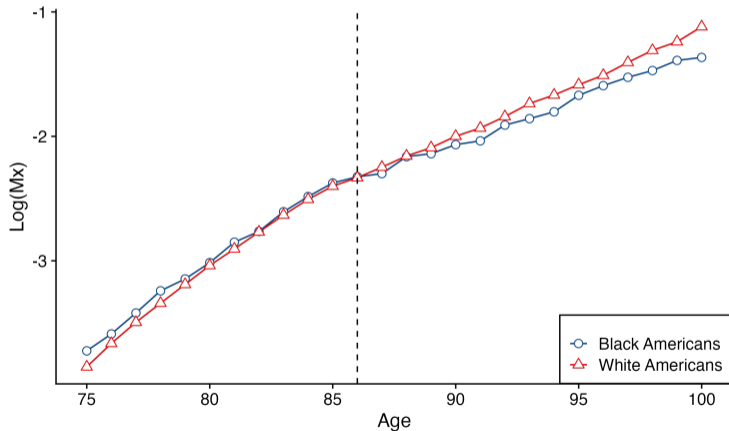
González et al. — Hispanic mortality paradox



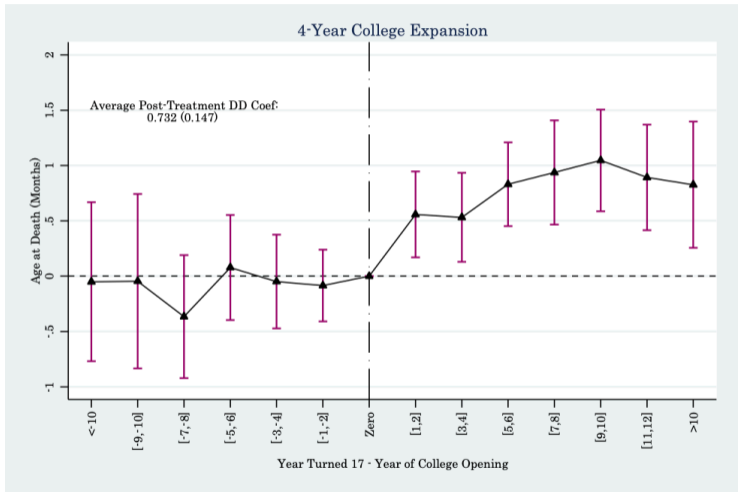
Black White Mortality Crossover

a Mortality Crossovers (Men)

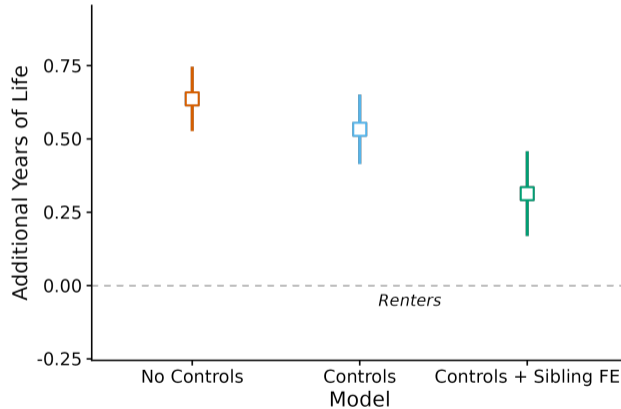
Pooled cohorts of 1890 - 1905



Fletcher et al. 2022 — Education Expansion



Longevity Benefits of Homeownership



Other Data Products

- ▶ Berkeley Unified Numident Mortality Database (BUNMD)

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

Other Data Products

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

- ▶ Berkeley Unified Numident Mortality Database (BUNMD)
 - ▶ 49 million death records
 - ▶ Date of birth, date of death, birthplace, race, sex, parents names, etc.

Other Data Products

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

- ▶ Berkeley Unified Numident Mortality Database (BUNMD)
 - ▶ 49 million death records
 - ▶ Date of birth, date of death, birthplace, race, sex, parents names, etc.
- ▶ World War II Army Enlistment Records
 - ▶ 9 million records, height + weight
 - ▶ Linked 1940 Census, linked mortality records

Other Data Products

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

- ▶ Berkeley Unified Numident Mortality Database (BUNMD)
 - ▶ 49 million death records
 - ▶ Date of birth, date of death, birthplace, race, sex, parents names, etc.
- ▶ World War II Army Enlistment Records
 - ▶ 9 million records, height + weight
 - ▶ Linked 1940 Census, linked mortality records

Considerations and future directions

- ▶ 1950 Census linkage

Considerations and future directions

- ▶ 1950 Census linkage
- ▶ Linkage to more recent death records (depending on quality)

Considerations and future directions

- ▶ 1950 Census linkage
- ▶ Linkage to more recent death records (depending on quality)
- ▶ **More substantive research mortality research**

Considerations and future directions

- ▶ 1950 Census linkage
- ▶ Linkage to more recent death records (depending on quality)
- ▶ **More substantive research mortality research**
 - ▶ Incorporate links into other linkage efforts (IPUMS MLP)

Thank You

Download: CenSoc.Berkeley.edu

Funding: R01AG058940, R01AG076830

Contact: ✉ casey.breen@demography.ox.ac.uk



scientific **data**



OPEN

DATA DESCRIPTOR

CenSoc: Public Linked Administrative Mortality Records for Individual-level Research

Casey F. Breen^{1,2}✉, Maria Osborne¹ & Joshua R. Goldstein¹✉

In the United States, much has been learned about the determinants of longevity from survey data and aggregated tabulations. However, the lack of large-scale, individual-level administrative mortality records has proven to be a barrier to further progress. We introduce the CenSoc datasets, which link the complete-count 1940 U.S. Census to Social Security mortality records. These datasets—CenSoc-DMF (N = 4.7 million) and CenSoc-Numident (N = 7.0 million)—primarily cover deaths among individuals aged 65 and older. The size and richness of CenSoc allows investigators to make new discoveries into geographic, racial, and class-based disparities in old-age mortality in the United States. This article gives an overview of the technical steps taken to construct these datasets, validates them using external aggregate mortality data, and discusses best practices for working with these datasets. The CenSoc datasets are publicly available, enabling new avenues of research into the determinants of mortality disparities in the United States.

Reserve Slides

Introduction
○○○○○○○○○

Creating CenSoc
○○○○○○○○○

Mortality Estimation
○○○○○○○

Case Studies
○○○○○

Conclusion
○○○

Reserve slides
●○

References

Goldstein et al. — Black names and longevity

Dependent Variable:	Death Age			
	Pooled	Family FE		
Model:	(1)	(3)	(4)	(5)
BNI (Standardized)	-0.2386 (0.2301)	-0.6258* (0.3060)	-0.6273* (0.3055)	-0.4696 (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:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

References

- 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. René 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.
- 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.