Using Machine Learning Algorithms to Predict Age of Death

Casey F. Breen[1] **Nathan Seltzer**[2]

[1, 2] Department of Demography | University of California, Berkeley

Authors contributed to this work equally

Social Science is increasingly interested in individual-level outcomes

- Researchers are increasingly seeking to pose and answer research questions about prediction at the individual-level (e.g., Hofman et al. 2017, Salganik et al. 2020, Arpino et al. 2022)
  - Increasing availability of rich individual-level data (digitization of census data, digital trace data, national register data, etc.)
- However, demographers still know relatively little about how accurately demographic events – such as fertility, migration, or mortality - can be predicted at the individual level

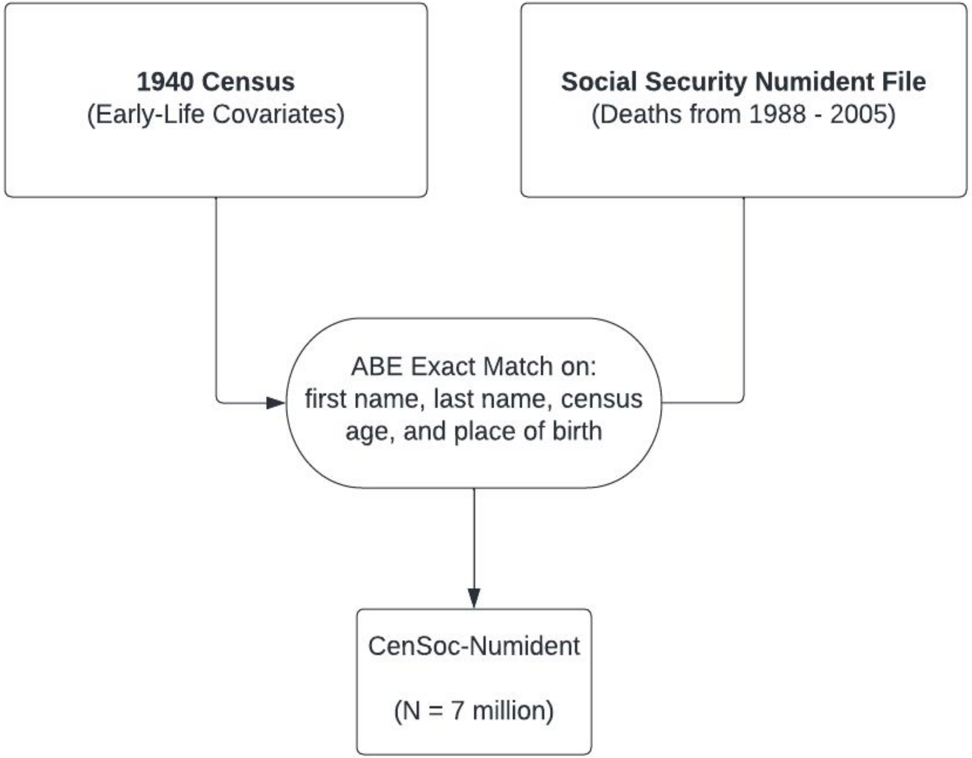
Research Question: How accurately can age of death be predicted from sociodemographic characteristics?

Answer speaks to the social rigidity of mortality: is human longevity a deterministic or stochastic process?

Answer to question has applications to:

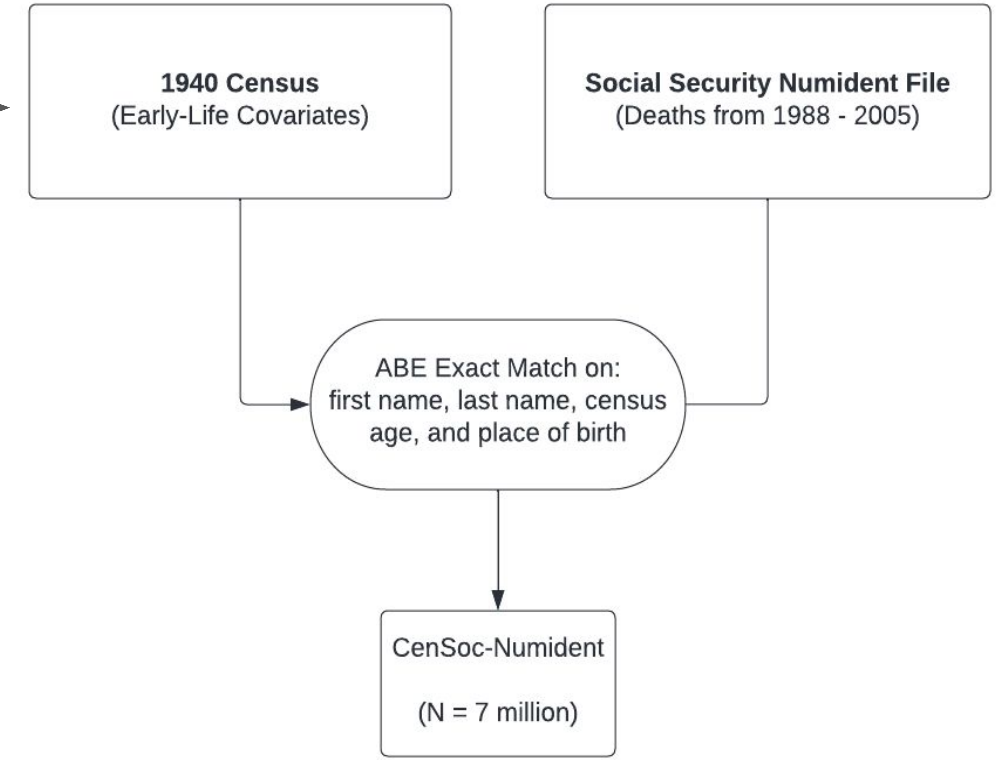
- Individual-level mortality risk scores used in medicine and epidemiology, where such mortality risk scores are valuable for adjusting for risk between treatment groups in both clinical and/or observational studies
- Mortality risk models could allow for more efficiently targeted individual-level treatments and interventions

CenSoc dataset: 1940 Census + Numident Mortality Records



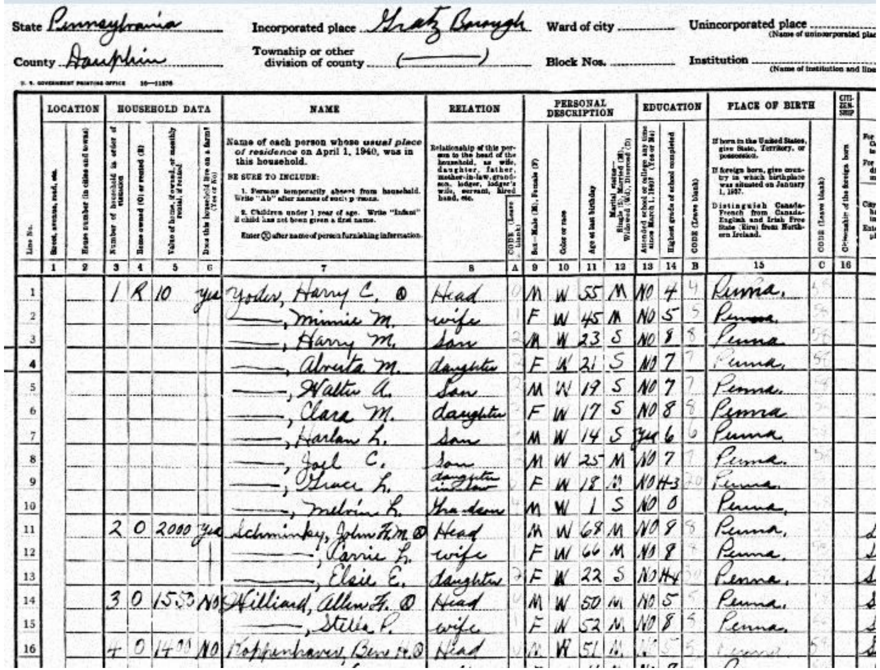
CenSoc dataset: 1940 Census + Numident Mortality Records

1940 Census is the first census to include questions about income, education, housing, etc.



Information collected on the 1940 Census

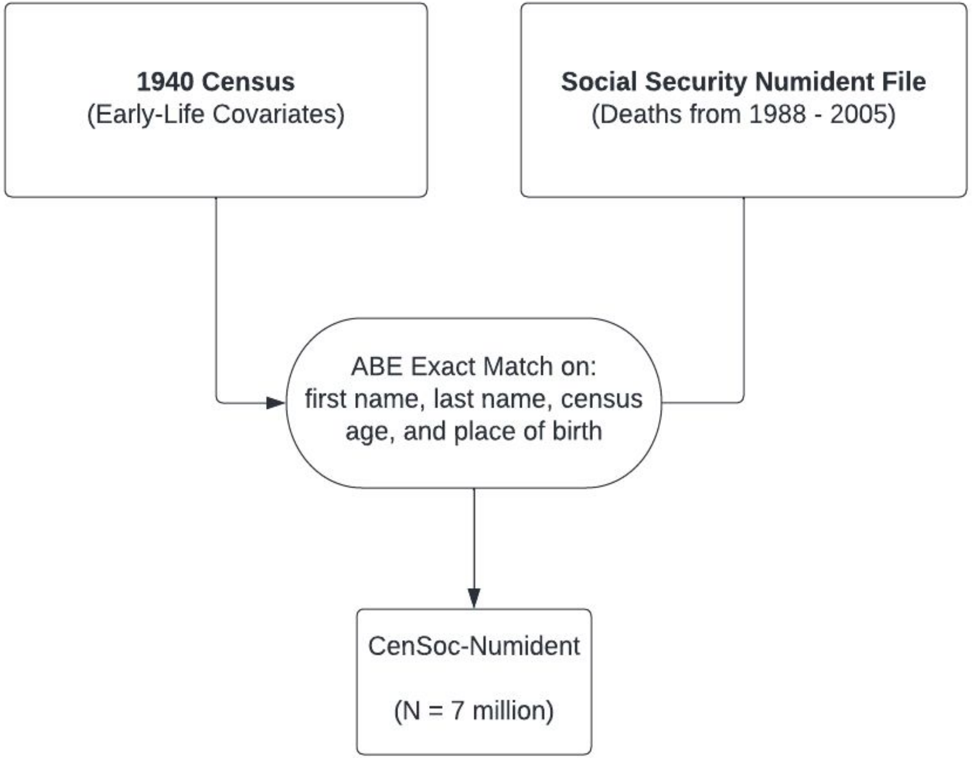
- Census Form:
  - Gender
  - Race
  - Place of birth
  - Internal migration
  - Age
  - Employment status / occupation*
  - Household characteristics
  - Education
  - Wage income*



State Pennsylvania — Incorporated place Gratz Borough — Ward of city — Unincorporated place
County Dauphin — Township or other division of county — Block Nos. — Institution

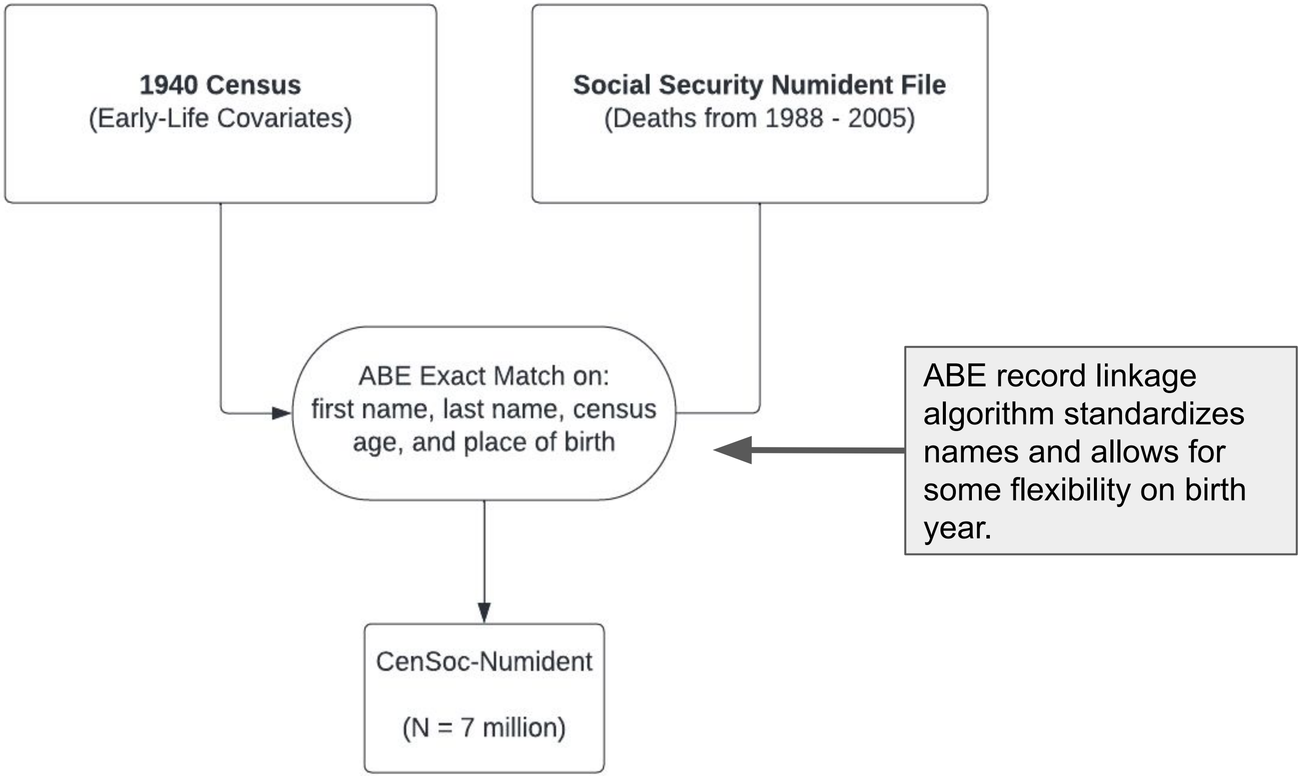
Line No.	House owned (O) or rented (R)	Value of home or monthly rental	Farm?	NAME	RELATION	Sex	Color or race	Age	Marital status	School	Highest grade	PLACE OF BIRTH
1	1 R	10	yes	Yoder, Harry C. ⊗	Head	M	W	55	M	No	4	Penna.
2				—, Minnie M.	wife	F	W	45	M	No	5	Penna.
3				—, Harry M.	Son	M	W	23	S	No	8	Penna.
4				—, Alverta M.	daughter	F	W	21	S	No	7	Penna.
5				—, Walter A.	Son	M	W	19	S	No	7	Penna.
6				—, Clara M.	daughter	F	W	17	S	No	8	Penna.
7				—, Harlan L.	Son	M	W	14	S	Yes	6	Penna.
8				—, Joel C.	Son	M	W	25	M	No	7	Penna.
9				—, Grace L.	daughter-in-law	F	W	18	M	No	H3	Penna.
10				—, Melvin L.	Grandson	M	W	1	S	No	0	Penna.
11	2 O	2000	yes	Schminky, John H.M. ⊗	Head	M	W	68	M	No	8	Penna.
12				—, Carrie L.	wife	F	W	66	M	No	8	Penna.
13				—, Elsie E.	daughter	F	W	22	S	No	H4	Penna.
14	3 O	1500	No	Hilliard, Allen H. ⊗	Head	M	W	50	M	No	5	Penna.
15				—, Stella P.	wife	F	W	52	M	No	8	Penna.
16	4 O	1400	No	Koppenhaver, Ben H. ⊗	Head	M	W	51	M	[illegible]	[illegible]	[illegible]

CenSoc dataset: 1940 Census + Numident Mortality Records



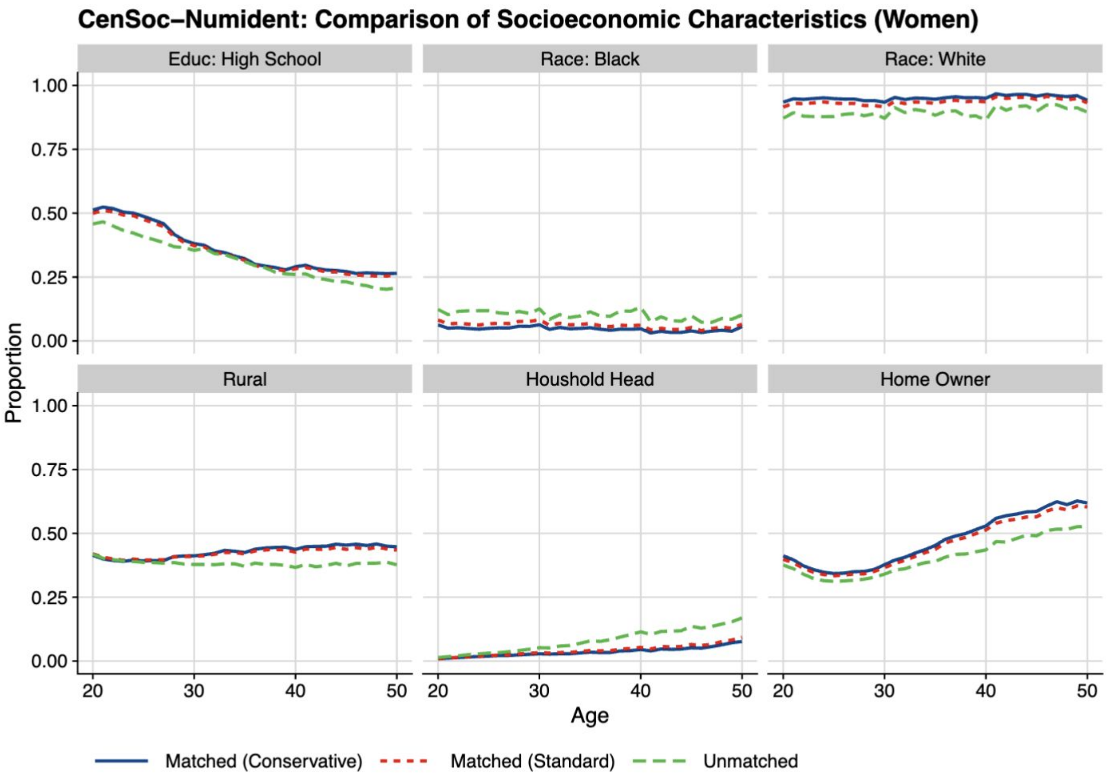
Publicly available mortality records, almost no sociodemographic characteristics.

CenSoc dataset: 1940 Census + Numident Mortality Records

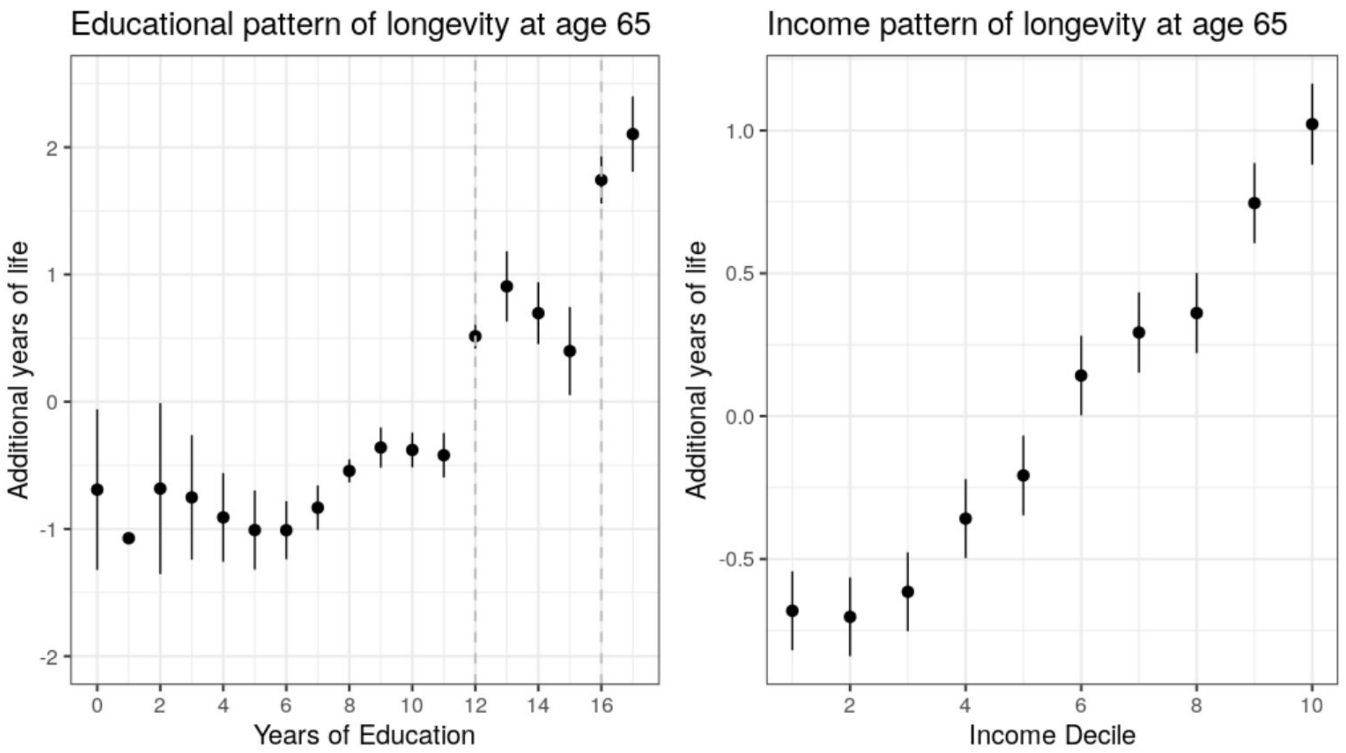


Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2019. Automated Linking of Historical Data.

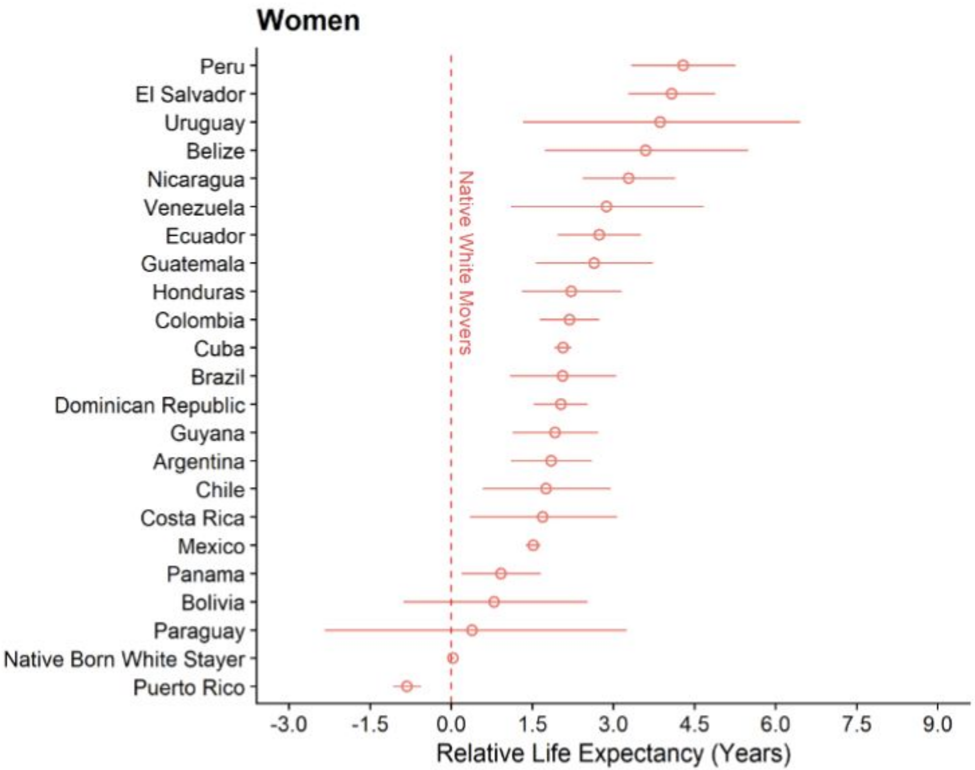
CenSoc-Numident is broadly representative of the population but contains slightly higher SES individuals



CenSoc allow us to zoom in on “high-resolution” aggregate mortality disparities (e.g., education staircase)



Clear country-of-origin patterns of longevity (Hispanic Mortality Paradox)



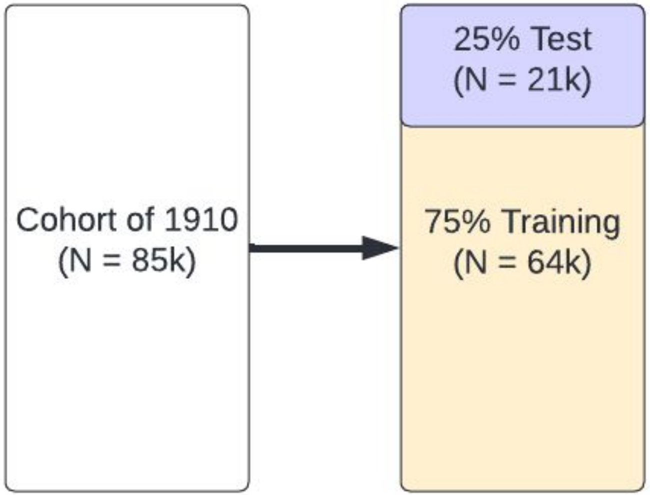
Andrea Miranda-Gonzalez, Kathy Perez, and Casey Breen. Understanding the Hispanic Mortality Paradox: Variation by Country of Origin

Can we predict later-life longevity using early-life sociodemographic characteristics?

Analytic Strategy

- Machine learning: Allows for detection of interaction terms and higher order effects
  - Primarily interested in prediction, not interpretability
- Train machine learning algorithms on randomly sampled “training” partition, test algorithms on the randomly sampled “testing” partition
- Restrict to cohort of 1910
  - Age 30 when observed in 1940 Census
  - Computationally easier, still large sample
  - Similar results for other cohorts + pooled cohorts
- Standardized continuous variables using min-max normalization

Sample Split

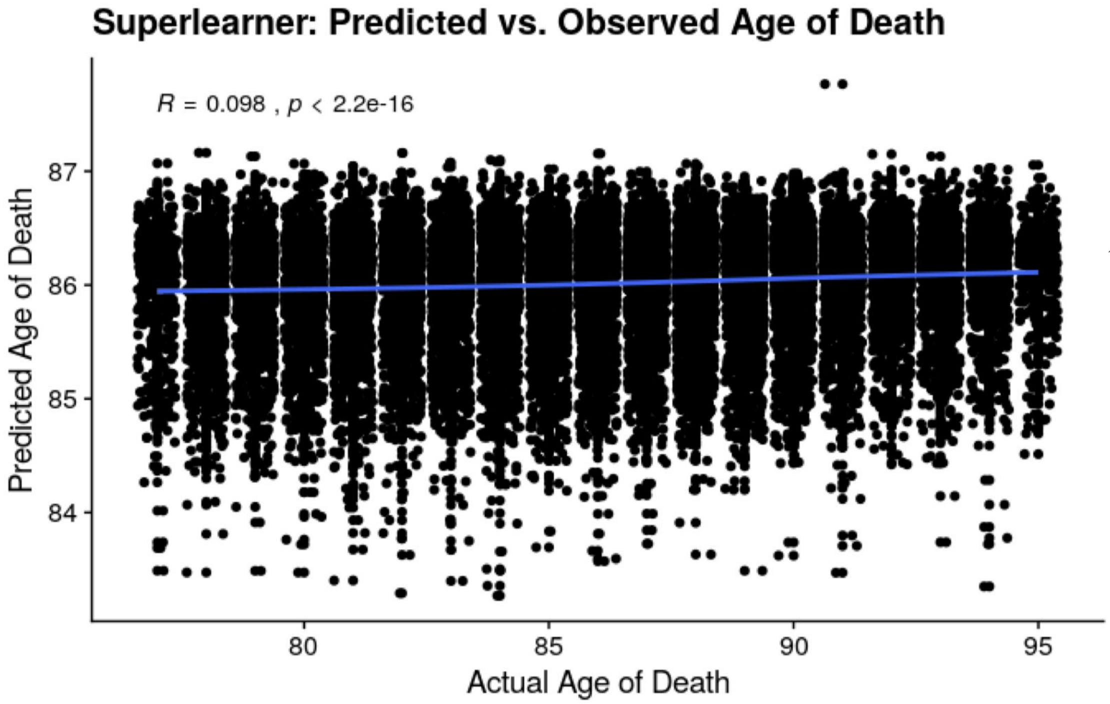


Superlearner — an ensembling approach

- How do you pick best machine learning algorithm?
- Superlearner (ensemble learning) fits several different algorithms and tests performance using cross-validation to estimate mean squared error for each algorithm
    - Combines models into a single model by picking the weighted combination of algorithms that has the lowest mean squared error

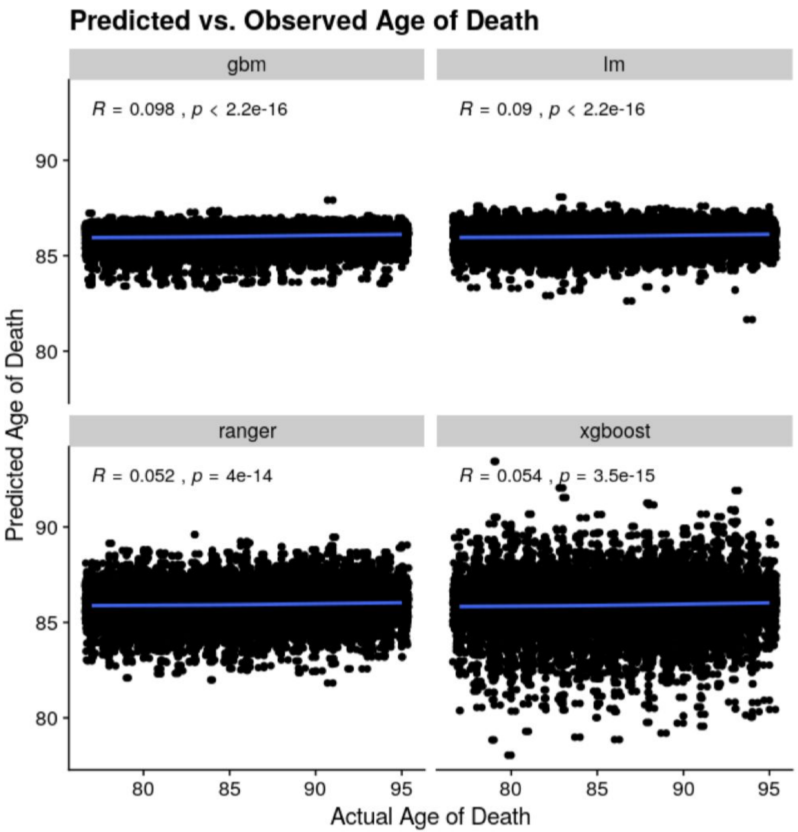
Algorithm	Description	Cross-validated Risk	Superlearner Coefficient
gbm	Generalized boosted regression	22.84	0.74
lm	Linear model	22.87	0.21
xgboost	Extreme gradient boosting	23.30	0.05
ranger	Random forest regression + classification	23.35	0.0001
ridge	Ridge Regression	22.87	0.0
mean	Arithmetic mean	23.11	0.0

Our best model explains ~1% of variation in age of death in testing dataset

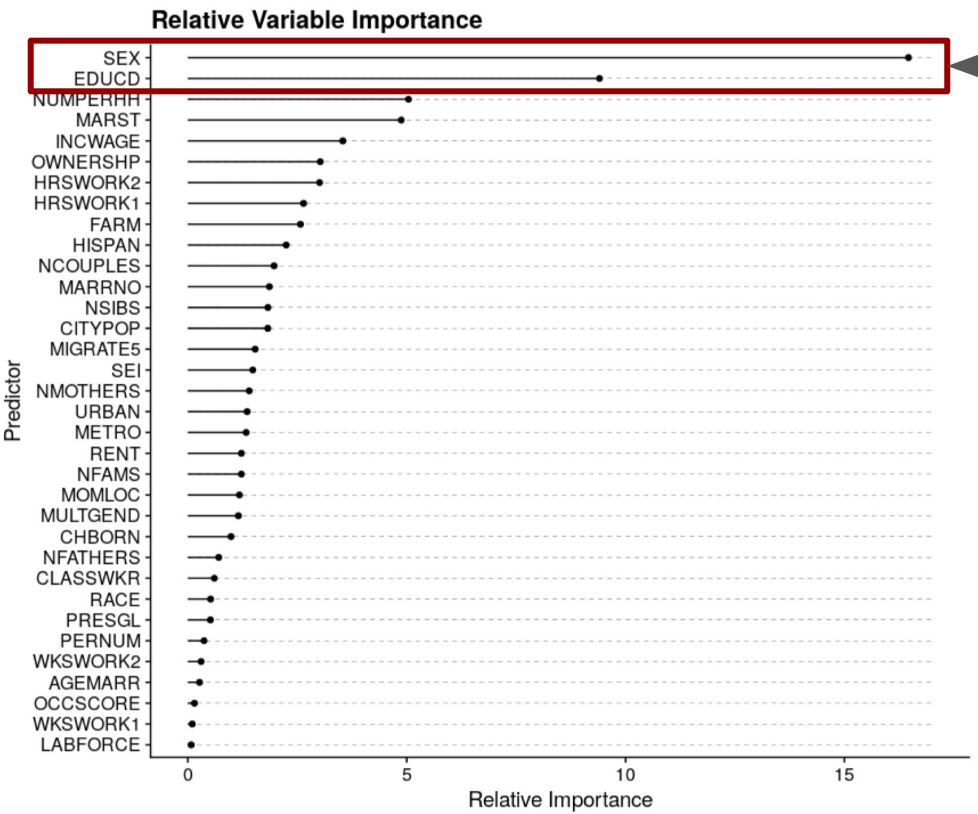


Low predictive accuracy ($R = .098$, $R^2 = .0096$).

Similar patterns across all machine learning algorithms tested

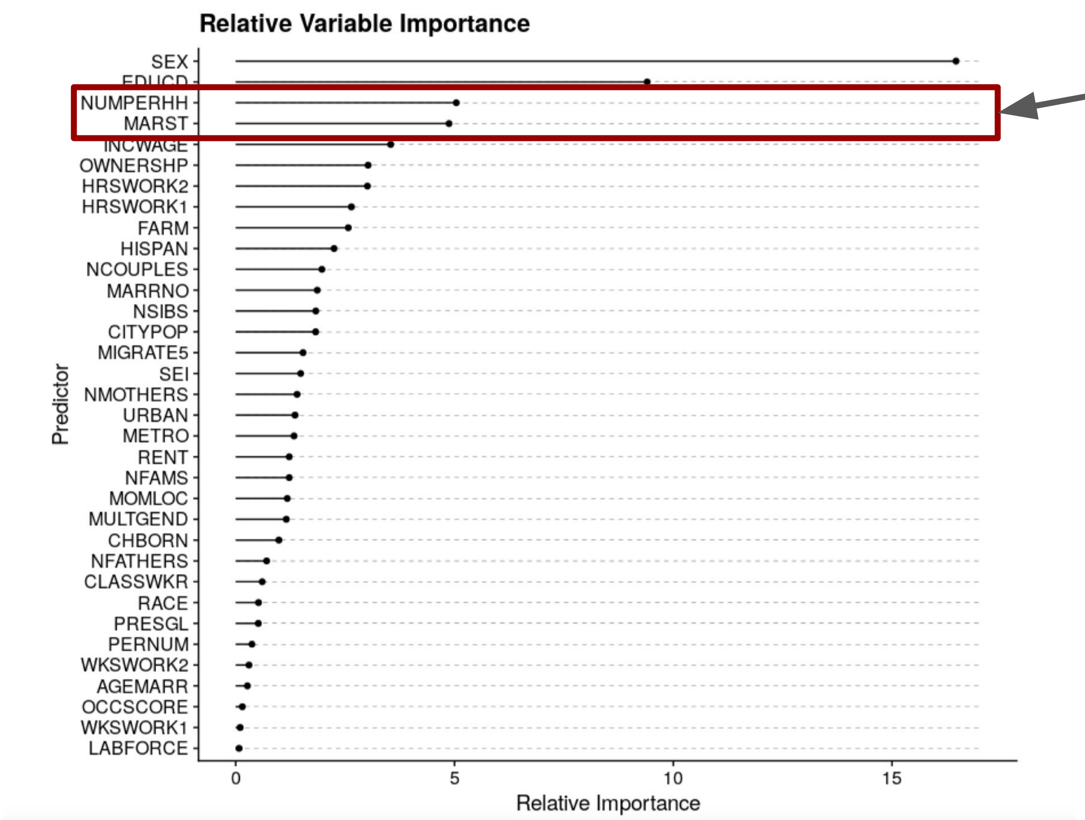


Gender and education are unsuprisingly the most ‘important’ predictors



Gender and education (in years) are the most important predictors

Household size and marital status are other key predictors



Number of persons living in the household and marital status are key predictors

Conclusions

- We fit several different machine learning algorithms on a large-scale mortality dataset, finding none of our algorithms predicted well in our test dataset
- Early life sociodemographic characteristics are very weak predictors of later life age of death
    - Even if we can see clear mortality disparities at the fine aggregate levels, this doesn't translate into the ability to predict individual-level outcomes
- Mortality is a stochastic process that isn't pre-determined: huge amounts of unobserved heterogeneity not captured by early-life sociodemographic characteristics

Thank you

Questions?

caseyfbreen

caseybreen@berkeley.edu