Towards a New, Public Dataset for Studying Mortality Inequality

Matching the 1940 U.S. Census with Social Security death records, 1975-2005

Joshua R. Goldstein Monica Alexander

UC Berkeley, Dept. of Demography

PAA Annual Meetings, Chicago
April 2017

Two quotations

The Human Mortality Database has launched thousands of papers, but we're at risk of falling behind. It's hard to study inequality without individual level data.
– PAA HMD workshop

Two quotations

The Human Mortality Database has launched thousands of papers, but we're at risk of falling behind. It's hard to study inequality without individual level data.
– PAA HMD workshop

We lost our access to the SIPP & SSA data file when the grant expired.
– a real researcher working on inequality and mortality

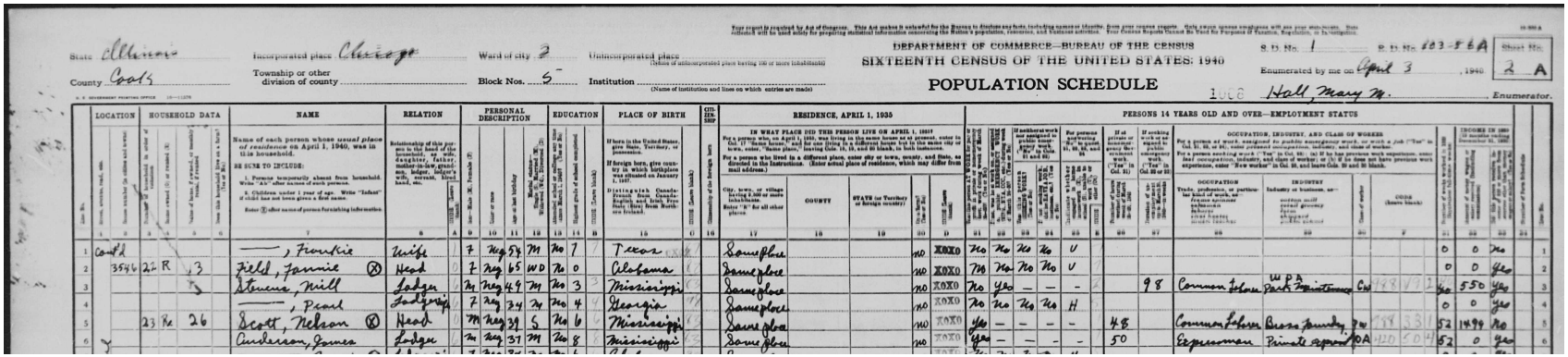
The Challenges

- Protect individual privacy
- Easy access for all researchers
- Replication and scientific progress
  (improving on others' work)
- Large sample sizes
- Lots of covariates
- Integrated estimation (matching and modeling)

The Opportunity: Unrestricted data

- The 1940 U.S. Census with names and rich covariates
- Social Security Death Records with names and mortality

1940 Census

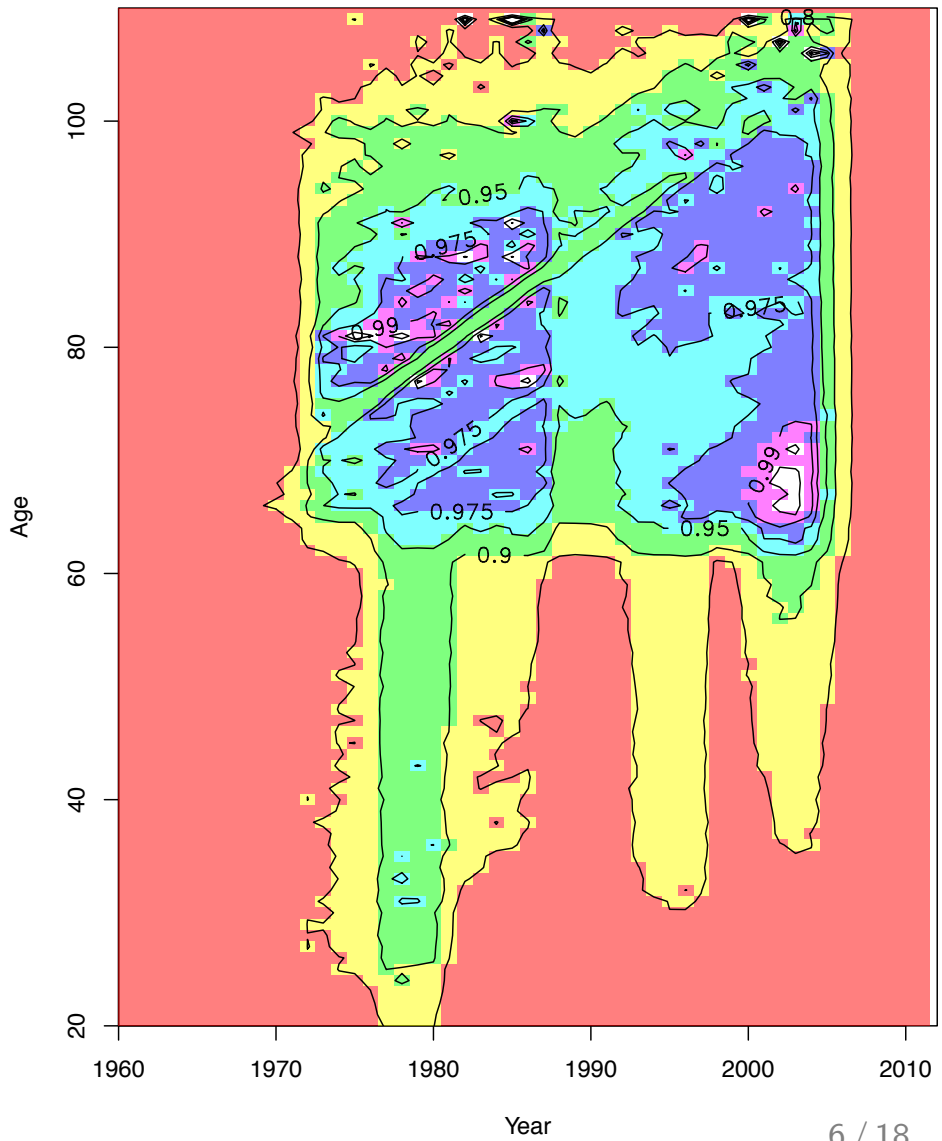


- 130 million people
- Rich individual level variables
  (Full demographics & income & education & housing)
- Street address level geography (Neighbors, blocks, ...)
- Already transcribed by Ancestry and MPC

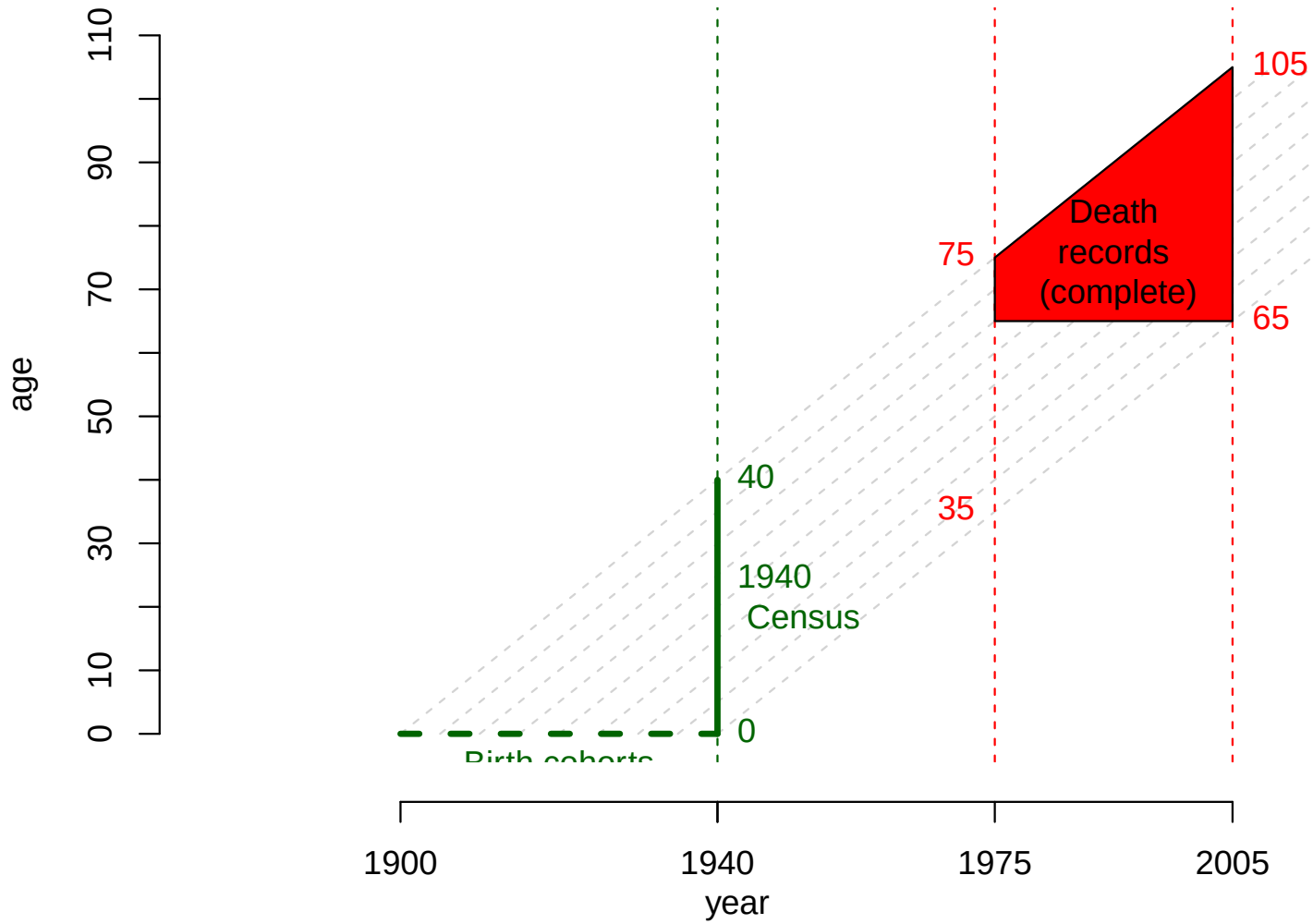
Social Security Death Index

- 80 million deaths
- Nearly complete coverage (over age 65, 1975-2005)
- Name, SSN, DOB, DOD
- Public information

SSDI deaths / HMD counts



Lexis diagram of CenSoc linkage



Matching method

Exact matching of unique keys

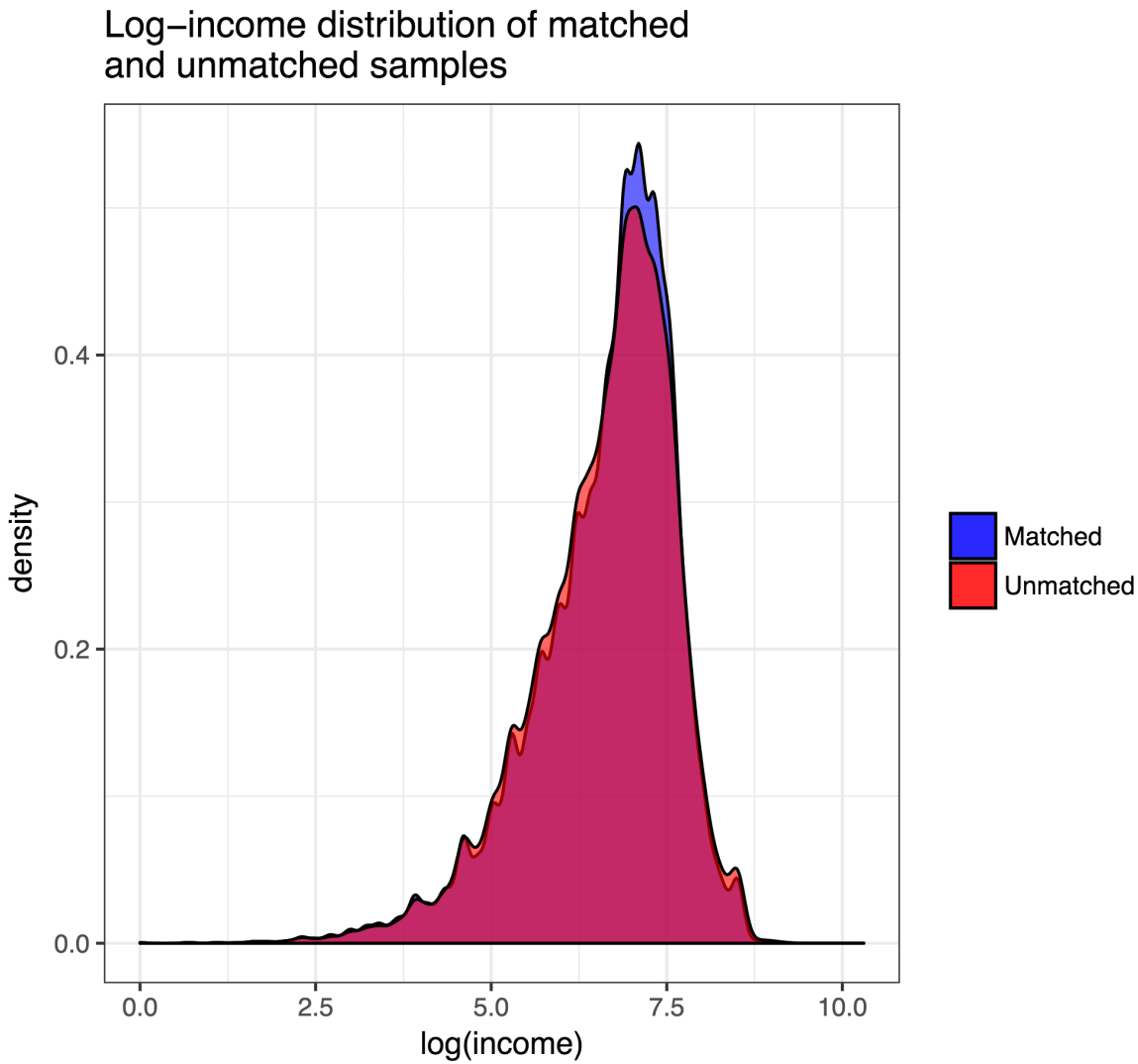
(first name, last name, birth year)

- About 75% of keys are unique

Linkage

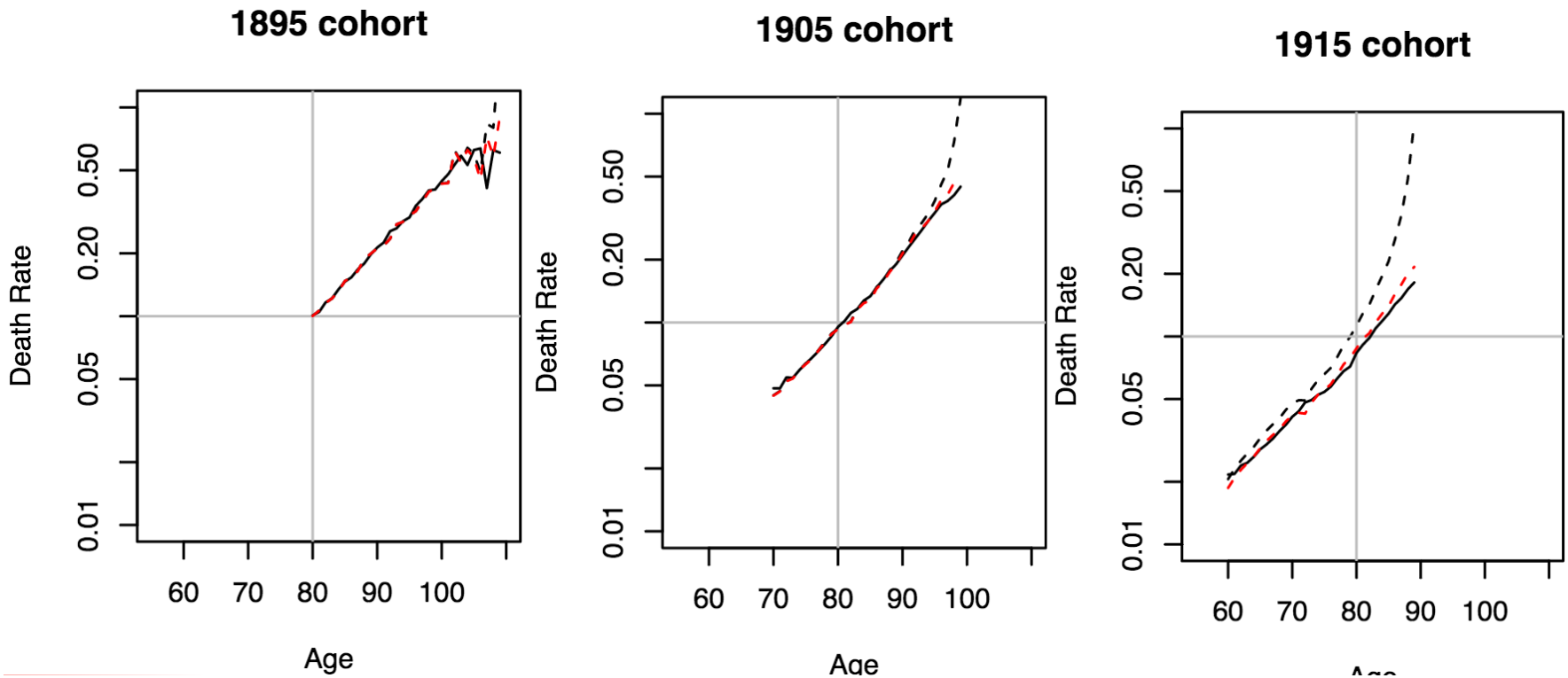
Full 1940 census	132	million
and name and age available	130	million
and male	64	million
and aged 0-70	62	million
and unique key	46	million
and expected death in interval	14	million
Matches	6	million
Match rate	43	%

Matching bias?

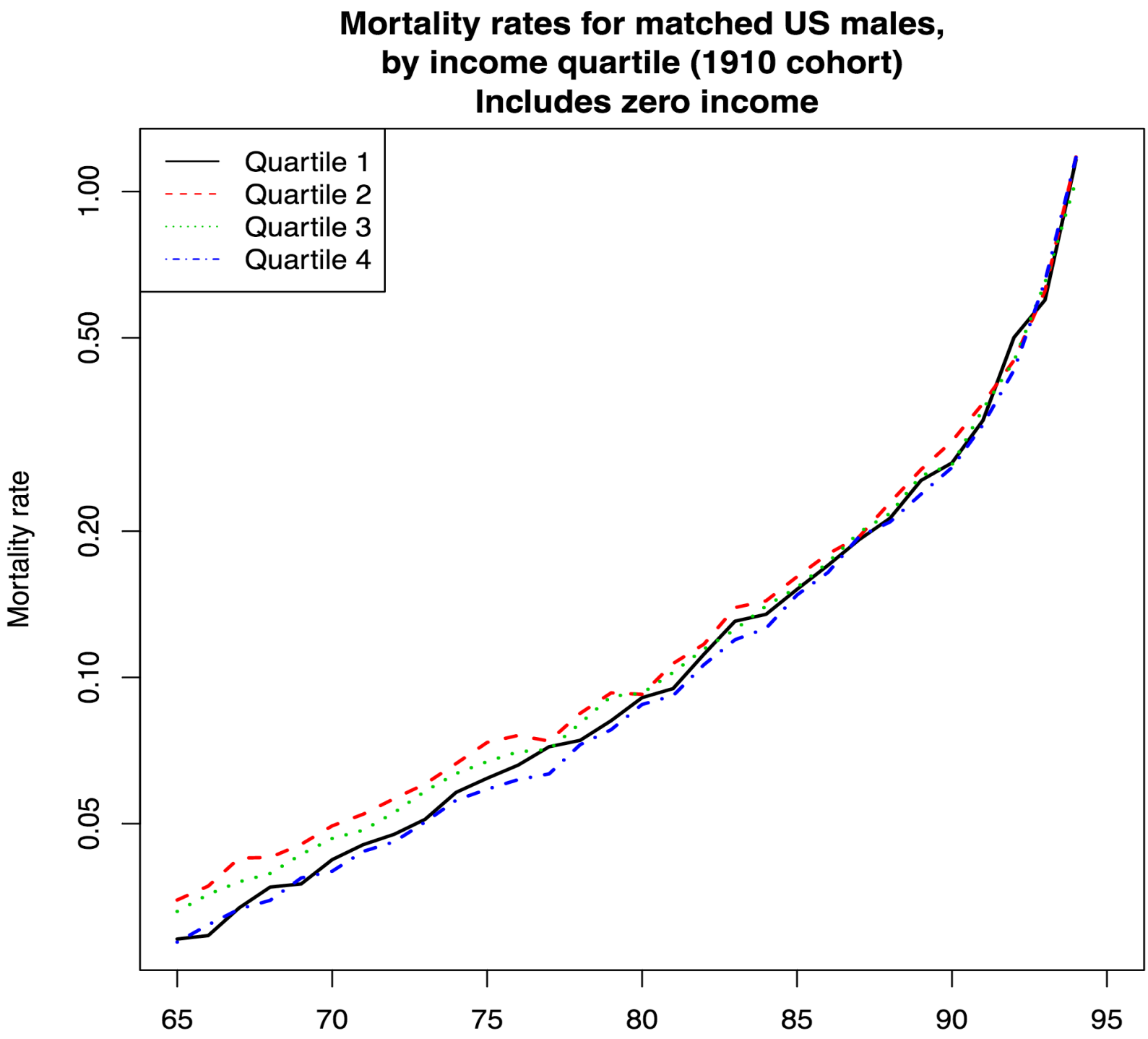


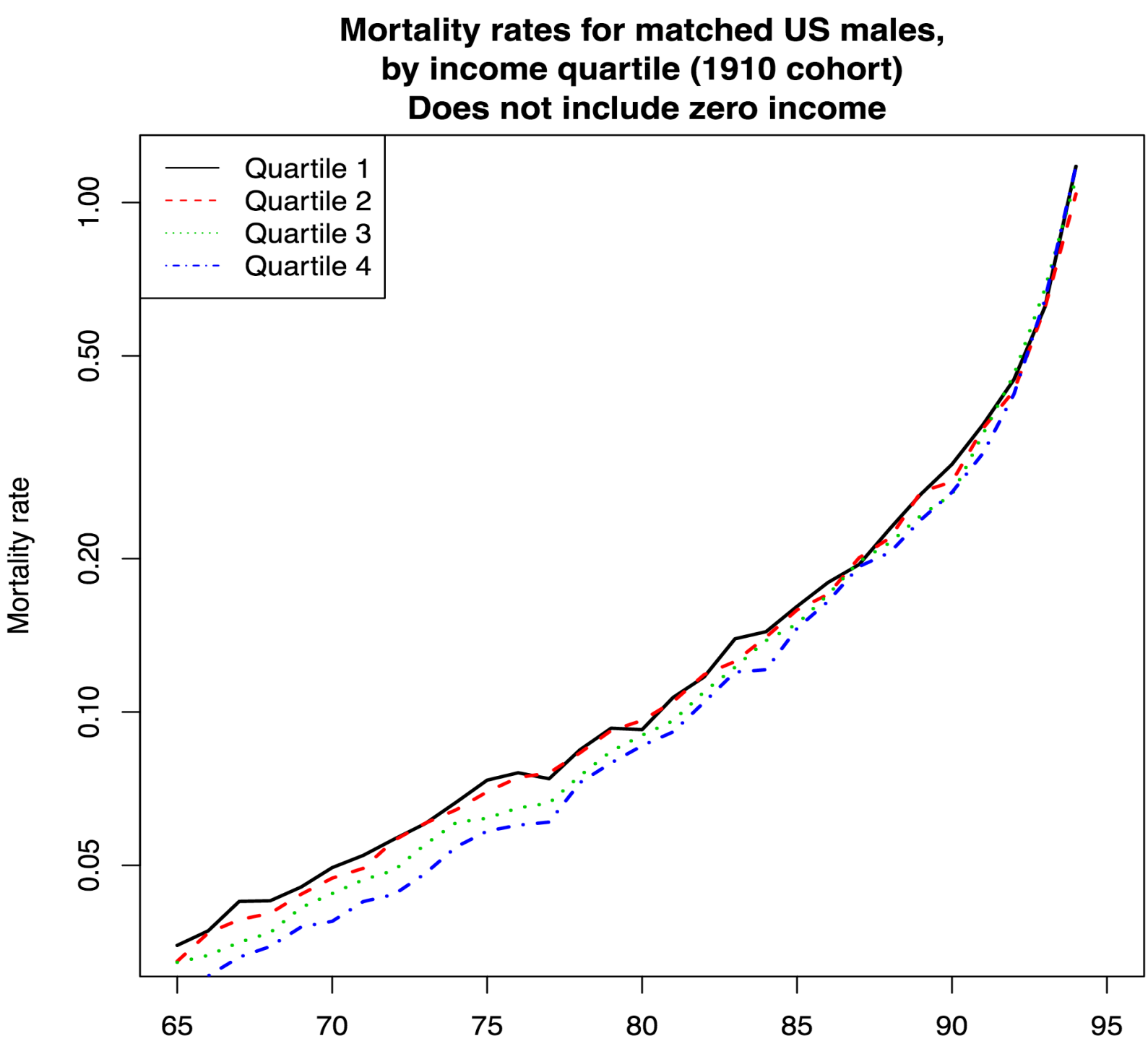
Matched also a bit whiter, more educated, and likely to be home-owners.

Mortality rate validation



Key:
solid line = HMD (Human Mortality Database)
black dash = reverse survival matched data
red dash = reverse survival matched data, adjusted with HMD survivors


Mortality rates for matched US males,
by income quartile (1910 cohort)
Includes zero income
Quartile 1
Quartile 2
Quartile 3
Quartile 4
Mortality rate
0.05
0.10
0.20
0.50
1.00
65
70
75
80
85
90
95


Mortality rates for matched US males,
by income quartile (1910 cohort)
Does not include zero income
Quartile 1
Quartile 2
Quartile 3
Quartile 4
Mortality rate
0.05
0.10
0.20
0.50
1.00
65
70
75
80
85
90
95

Smaller groups and new variables

OLS regressions of age-at-death for those aged 20-35 in 1940

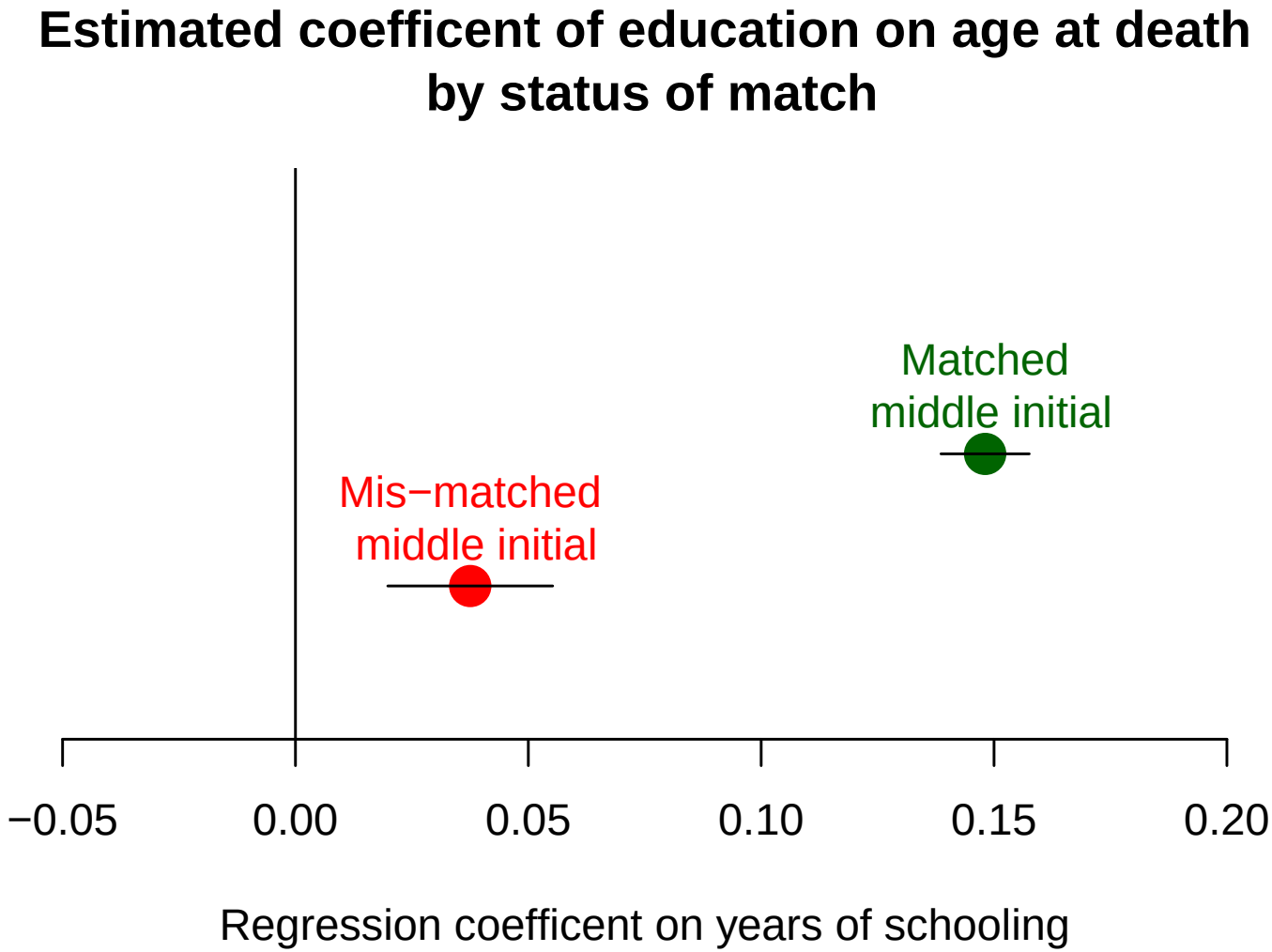
Intercept	77.83***	74.70***
Black/White	-0.90***	0.08
Chinese/White	1.01***	1.90***
Filipino/White	1.69***	2.40***
Japanese/White	1.90***	2.26***
Other/White	-1.27***	-0.70***
educ		0.19***
log(income)		0.18***
own/rent		0.50***
hh_head: Yes		-0.15
N	2 million	1.1 million

Validation: Middle Initials

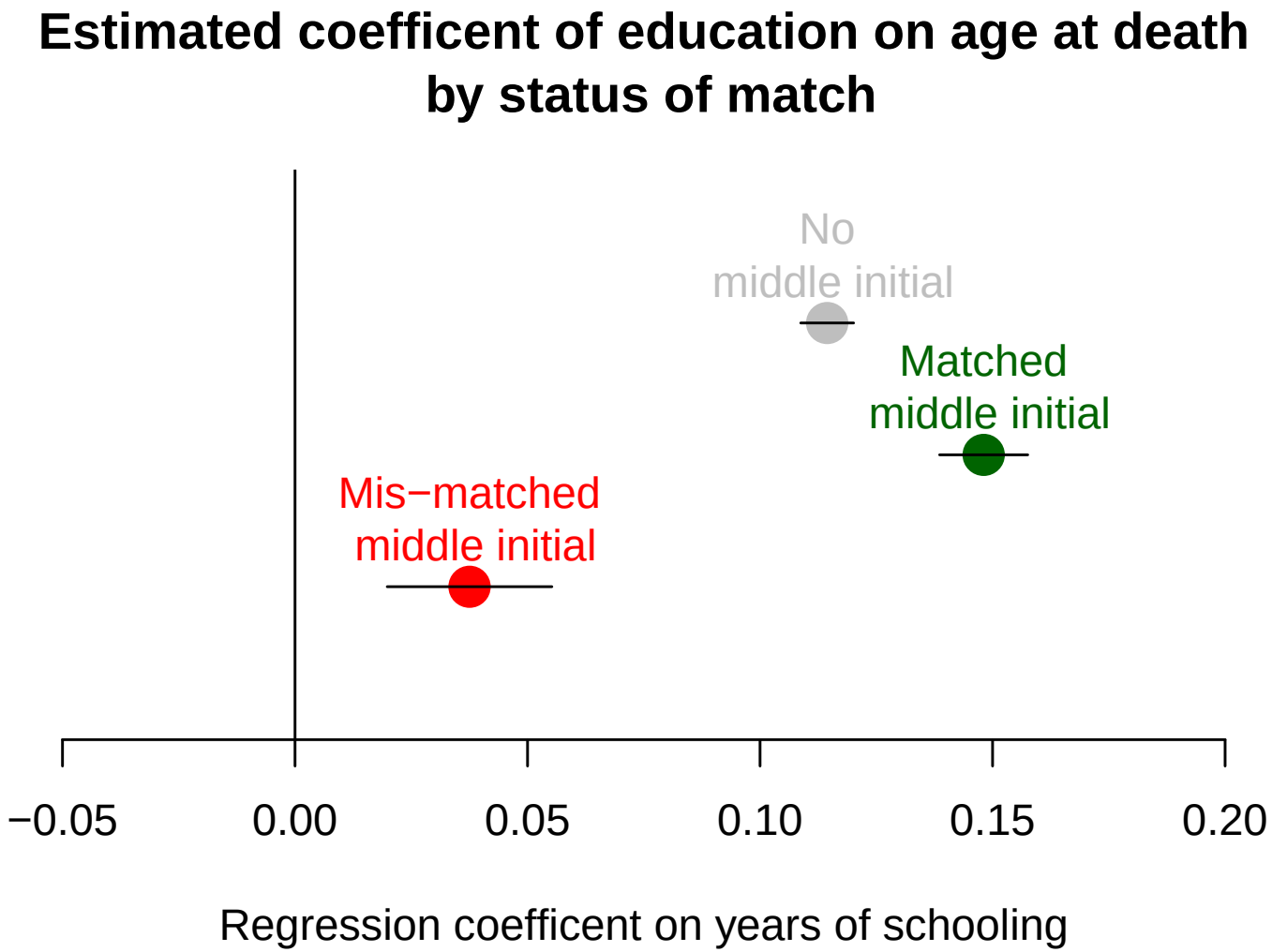
	clean_key	ssn	fname.x	lname.x	mname.x	mname	mi.match
1:	AABERGEDWARD49	538074776	EDWARD	AABERG	J		
2:	AABERGEELMER32	516228997	ELMER	AABERGE			
3:	AABERGERIC34	521071090	ERIC	AABERG	C	C	TRUE
4:	AABERGLAWRENCE22	517169163	LAWRENCE	AABERG	M	A	FALSE
5:	AABERGRALPH30	522071496	RALPH	AABERG		O	
6:	AABERGROBERT39	563033374	ROBERT	AABERG	A	A	TRUE
7:	AABERGSANDER43	535096685	SANDER	AABERG	P		
8:	AABWILLIAM42	523147290	WILLIAM	AAB			
9:	AABYCARLYLE20	473091698	CARLYLE	AABY	P	P	TRUE
10:	AABYELWIN18	517167623	ELWIN	AABY			

Middle initial match rate $\approx 80\%$

Validation: Errors-in-variables framework



Validation: Errors-in-variables framework



Future Directions

- Public release of our 6-7 million linked deaths (with IPUMS, HMD)
- Other match methods (birthplace, probablistic matches?)
- Mortality estimation for linked data
  - Parametric MLE for doubly-truncated cohorts
  - Bayesian methods (Schmertmann et al.)
  - Missing-at-random methods (Taylor, Sanders et al.)